Limpieza y transformación de datos: técnicas y herramientas informáticas

Breve descripción:

Este componente presenta una exploración de los procesos de limpieza y transformación de datos, abarcando desde sus fundamentos hasta las técnicas y herramientas más recientes.

Noviembre 2024

Tabla de contenido

Introducción

En el mundo actual del análisis de datos, la calidad y la preparación adecuada de la información se han convertido en factores determinantes para el éxito de cualquier proyecto analítico. Los datos crudos, aquellos que provienen directamente de diversas fuentes, raramente se encuentran en un estado óptimo para su análisis inmediato. Frecuentemente, contienen inconsistencias, valores faltantes, formatos incompatibles y otros problemas que pueden comprometer la validez de los análisis posteriores.

Pero ¿cómo se transforma un conjunto de datos desordenado en información valiosa y procesable? Este componente formativo aborda esta pregunta fundamental, adentrándose en el mundo de la limpieza y transformación de datos. Se exploran las técnicas, herramientas y mejores prácticas que permiten convertir datos crudos en conjuntos de información confiables y útiles para el análisis.

A lo largo de ese componente formativo, se guiará al aprendiz a través de un proceso sistemático que abarca desde la importación inicial de datos hasta su transformación final, pasando por técnicas de limpieza, validación y control de calidad. Mediante ejemplos prácticos y casos de estudio reales, se descubrirá cómo abordar los desafíos comunes en la preparación de datos y cómo implementar soluciones efectivas utilizando herramientas modernas como Python y R.

La limpieza y transformación de datos son pasos preliminares en el proceso analítico, pero también constituyen la base fundamental sobre la cual se construye todo análisis exitoso. Como reza el conocido adagio en ciencia de datos: “garbage in, garbage out” (si entra basura, sale basura). Por ello, es necesario dominar estas técnicas y herramientas, para cualquier persona que trabaje con datos.

¡Bienvenido a este recorrido por el mundo de la estadística y la visualización efectiva de datos para el análisis!

Video 1. Limpieza y transformación de datos técnicas y herramientas informáticas

Enlace de reproducción del video

Síntesis del video: Limpieza y transformación de datos técnicas y herramientas informáticas
En el componente formativo «Limpieza y transformación de datos: técnicas y herramientas informáticas» se explora el proceso de preparación de datos para el análisis, un aspecto fundamental en cualquier proyecto de analítica.

La calidad de los datos es determinante para el éxito de los análisis. Como dice el adagio: "garbage in, garbage out" (si entra basura, sale basura). Los científicos de datos dedican incluso el 80% de su tiempo a estas tareas de preparación.

Los datos pueden presentarse en tres formas principales: estructurados (como tablas ordenadas), no estructurados (como textos o imágenes), y semiestructurados (como archivos JSON). Cada tipo requiere técnicas específicas de procesamiento.

La importación de datos abarca desde la lectura de archivos simples hasta conexiones complejas con bases de datos. Las técnicas de validación y las mejores prácticas en esta etapa son muy importantes para garantizar la integridad de la información.

Las técnicas de transformación incluyen la conversión de tipos de datos, clasificación, eliminación de datos innecesarios y creación de nuevas variables. Estas operaciones son esenciales para preparar los datos para el análisis.

Herramientas como Python y R, junto con sus bibliotecas especializadas como Pandas y dplyr, facilitan estas tareas. La automatización de procesos mediante pipelines mejora la eficiencia y reduce errores.

El control de calidad y la documentación son aspectos estratégicos. La trazabilidad de las transformaciones y la validación continua aseguran la confiabilidad de los resultados.

Las tendencias apuntan hacia una mayor automatización inteligente y democratización de estas herramientas, permitiendo que más profesionales participen en el proceso de preparación de datos.

¡Bienvenidos al mundo de la limpieza y transformación de datos!

1. Fundamentos de la limpieza y transformación de datos

La preparación y limpieza de datos constituye una fase decisiva y fundamental en cualquier proceso de análisis de datos. En este capítulo se introducen los conceptos esenciales relacionados con la limpieza y transformación de datos, procesos que consumen la mayor parte del tiempo en proyectos de análisis y de los cuales depende significativamente la calidad de los resultados obtenidos. Se examinan los diferentes tipos de datos que se pueden encontrar, así como los fundamentos del proceso de preparación de datos, estableciendo una base para comprender las técnicas y herramientas que se abordan posteriormente. La comprensión de estos conceptos fundamentales resulta esencial para desarrollar procesos de análisis de datos efectivos y confiables.

1.1. Introducción a la limpieza y transformación de datos

En el contexto actual de la ciencia de datos y la analítica avanzada, la limpieza y transformación de datos constituyen procesos fundamentales que determinan la calidad y confiabilidad de cualquier análisis posterior. Estos procesos representan la base sobre la cual se construyen los modelos analíticos y se fundamentan las decisiones basadas en datos.

La limpieza de datos se define como el conjunto de procedimientos sistemáticos destinados a identificar, corregir o eliminar inconsistencias, errores y anomalías presentes en los conjuntos de datos. Este proceso abarca desde la detección de valores faltantes hasta la corrección de discrepancias en los formatos y la eliminación de duplicados.

Por su parte, la transformación de datos comprende las operaciones necesarias para convertir los datos desde su estado original a un formato que resulte más adecuado para el análisis. Este proceso puede incluir la normalización de variables, la creación de nuevas características derivadas, y la reestructuración de los datos para satisfacer los requisitos específicos de las herramientas analíticas.

1.2. Importancia en el análisis de datos

La relevancia de la limpieza y transformación de datos trasciende el simple procesamiento técnico, constituyéndose en un factor muy importante para el éxito de cualquier proyecto de análisis de datos. Se estima que los científicos de datos —personas que analizan, procesan, y modelan grandes cantidades de datos— dedican entre el 60% y el 80% de su tiempo a estas tareas, lo cual refleja su importancia fundamental en el ciclo de vida del análisis de datos.

La calidad de los datos impacta directamente en la validez de las conclusiones extraídas del análisis. Datos incorrectos o mal estructurados pueden conducir a interpretaciones erróneas y, consecuentemente, a decisiones equivocadas. La precisión y confiabilidad de los modelos predictivos, los análisis estadísticos y las visualizaciones dependen fundamentalmente de la calidad de los datos subyacentes.

Además, la limpieza y transformación adecuada de los datos contribuye a la eficiencia computacional. Datos bien estructurados y limpios requieren menos recursos de procesamiento y permiten la aplicación más efectiva de algoritmos analíticos. Este aspecto resulta particularmente relevante cuando se trabaja con grandes volúmenes de datos o cuando se requieren análisis en tiempo real.

1.3. Tipos de datos

En la era del Big Data, el mundo se encuentra inmerso en un océano de información que fluye constantemente desde una miríada de fuentes. Comprender los diferentes tipos de datos y sus orígenes es fundamental para cualquier persona que trabaje en el campo de la analítica de datos. A continuación, se exploran en detalle los principales tipos de datos y las diversas fuentes de las que provienen.

- **Datos estructurados**

  Los datos estructurados son quizás los más familiares para la mayoría de las personas. Se caracterizan por su organización clara y predefinida, lo que los hace fácilmente identificables y analizables. Un ejemplo serían una hoja de cálculo perfectamente organizada o una base de datos relacional bien diseñada. Cada pieza de información tiene su lugar asignado, como fichas en un archivador meticulosamente organizado.

  En el mundo de los negocios, los datos estructurados abundan. Los registros de ventas, la información de clientes en un CRM, los datos de inventario o los registros de transacciones financieras son ejemplos clásicos. Estos datos suelen almacenarse en bases de datos relacionales y se pueden consultar fácilmente utilizando SQL.

  La belleza de los datos estructurados radica en su simplicidad y eficiencia. Son ideales para análisis cuantitativos, generación de informes y tableros (dashboards). Sin embargo, su rigidez puede ser también una limitación. En un mundo que cambia rápidamente, la estructura predefinida puede volverse obsoleta o insuficiente para capturar nuevas formas de información.

- **Datos no estructurados.**

  En contraste con sus contrapartes estructuradas, los datos no estructurados son como el arte abstracto del mundo de los datos. No siguen un formato o modelo predefinido y vienen en una variedad casi infinita de formas. Textos de redes sociales, correos electrónicos, archivos de audio, videos, imágenes e incluso el contenido de sitios web entran en esta categoría.

  El auge de Internet y las redes sociales ha provocado una explosión de datos no estructurados. Cada tweet, cada publicación de blog, cada video de YouTube es una pieza de datos no estructurados. Estos datos son ricos en información y contexto, pero presentan desafíos significativos para su análisis.

  Trabajar con datos no estructurados requiere técnicas y herramientas especializadas. El procesamiento del lenguaje natural (NLP) se utiliza para extraer significado de textos, mientras que el reconocimiento de imágenes y el procesamiento de señales se aplican a datos visuales y de audio.

  Aunque el análisis de datos no estructurados puede ser complejo, ofrece insights invaluables que los datos estructurados por sí solos no pueden proporcionar.

- **Datos semiestructurados.**

  Entre los mundos ordenados de los datos estructurados y el caos creativo de los no estructurados, se encuentran los datos semiestructurados. Estos datos tienen algún nivel de organización, pero no se ajustan perfectamente al modelo rígido de las bases de datos relacionales.

Los formatos JSON y XML son ejemplos clásicos de datos semiestructurados. Estos formatos proporcionan una estructura flexible que permite representar información compleja y jerárquica. Los documentos JSON, por ejemplo, son ampliamente utilizados en aplicaciones web y móviles para intercambiar datos.

Los datos semiestructurados ofrecen un equilibrio entre flexibilidad y organización. Son lo suficientemente flexibles como para adaptarse a cambios en la estructura de la información, pero mantienen suficiente orden como para facilitar su procesamiento y análisis. Las bases de datos NoSQL, como MongoDB, están diseñadas específicamente para manejar este tipo de datos de manera eficiente.

- **Fuentes de datos.**

  Las fuentes de datos en el ecosistema del Big Data son tan diversas como los datos mismos. Se pueden categorizar en tres grandes grupos: fuentes públicas, privadas y mixtas.

  Las fuentes de datos públicos son un tesoro de información accesible para cualquiera. Gobiernos de todo el mundo están adoptando políticas de datos abiertos, publicando conjuntos de datos sobre una amplia gama de temas, desde estadísticas demográficas hasta datos meteorológicos y ambientales. Organizaciones internacionales como las Naciones Unidas y el Banco Mundial también proporcionan vastos repositorios de datos globales. Además, muchas instituciones académicas y de investigación comparten sus datos para fomentar la colaboración y el avance científico.

Las redes sociales y plataformas en línea son otra fuente rica de datos públicos. Twitter, por ejemplo, ofrece acceso a su API, permitiendo a los investigadores y analistas estudiar tendencias y opiniones públicas en tiempo real. Sitios como Wikipedia no solo proporcionan contenido enciclopédico, sino también datos sobre patrones de edición y contribución de usuarios.

En el otro extremo del espectro están las fuentes de datos privados. Estos son los datos generados y mantenidos por empresas y organizaciones para su uso interno. Registros de clientes, datos de transacciones, información de empleados y datos operativos son ejemplos de datos privados. Estos datos suelen ser altamente valiosos y sensibles, y su acceso está estrictamente controlado.

Las empresas utilizan sus datos privados para impulsar la toma de decisiones, mejorar la eficiencia operativa y obtener ventajas competitivas. Sin embargo, el manejo de datos privados conlleva grandes responsabilidades, especialmente en lo que respecta a la privacidad y la seguridad. Regulaciones como el GDPR en Europa y el CCPA en California han establecido estrictos requisitos para el manejo de datos personales.

Las fuentes mixtas representan un área interesante donde los límites entre lo público y lo privado se difuminan. Un ejemplo son los datos generados por dispositivos IoT (Internet de las Cosas). Un termostato inteligente en un hogar genera datos privados sobre el uso de energía de una familia, pero cuando estos datos se agregan a nivel de ciudad o región, pueden convertirse en una valiosa fuente de información pública sobre patrones de consumo energético.

Otro ejemplo de fuente mixta son las plataformas de crowdsourcing y ciencia ciudadana. Proyectos como eBird, donde aficionados a las aves reportan sus avistamientos, crean conjuntos de datos que son a la vez personales y públicos, y que han demostrado ser invaluables para la investigación científica.

1.4. Proceso general de preparación de datos

El proceso de preparación de datos sigue una secuencia lógica de etapas interrelacionadas. Inicialmente, se realiza una evaluación preliminar de la calidad y estructura de los datos disponibles. Esta evaluación permite identificar los principales desafíos y determinar las estrategias más apropiadas para abordarlos.

La fase de limpieza incluye la identificación y tratamiento de valores atípicos, la gestión de valores faltantes, la eliminación de duplicados y la corrección de inconsistencias en los datos. Cada una de estas tareas requiere un análisis cuidadoso para determinar la acción más apropiada, considerando el contexto específico del problema y los requisitos del análisis posterior.

La transformación de datos puede incluir múltiples operaciones como la normalización de variables numéricas, la codificación de variables categóricas, la agregación de registros, y la creación de nuevas variables derivadas. La selección de las transformaciones específicas depende de los objetivos del análisis y de los requisitos de las técnicas analíticas que se planea utilizar.

La validación constituye una etapa fundamental del proceso, donde se verifica la integridad y coherencia de los datos procesados. Esta fase incluye la comprobación de restricciones lógicas, la verificación de rangos válidos para las variables, y la

confirmación de que las transformaciones realizadas han producido los resultados esperados.

El proceso de preparación de datos no debe considerarse como una secuencia lineal, sino como un proceso iterativo donde pueden requerirse múltiples ciclos de limpieza y transformación hasta alcanzar el nivel de calidad deseado. La documentación detallada de cada paso es fundamental para garantizar la reproducibilidad del proceso y facilitar futuras actualizaciones o modificaciones.

La figura que se presenta a continuación sintetiza los elementos centrales presentados en este apartado sobre la preparación de datos.

Figura 1. Proceso iterativo para obtener datos de calidad óptima

Fuente. OIT, 2024.

2. Importación y lectura de datos

La importación y lectura de datos representa una etapa obligada en el proceso de análisis, constituyendo el punto de entrada para toda la información que será procesada posteriormente. Este capítulo explora las diferentes técnicas y consideraciones necesarias para garantizar una importación de datos eficiente y confiable, abordando tanto la lectura desde archivos externos como la conexión con bases de datos.

2.1. Lectura de datos desde archivos externos

Los archivos CSV (valores separados por comas) constituyen uno de los formatos más utilizados en el intercambio de datos tabulares. Su popularidad radica en su simplicidad y universalidad, características que los hacen ideales para el intercambio de información entre diferentes sistemas y plataformas. Sin embargo, el trabajo con archivos CSV presenta desafíos particulares que deben abordarse cuidadosamente. La detección correcta de delimitadores, el manejo de caracteres especiales y la interpretación adecuada de tipos de datos son aspectos críticos que requieren atención detallada durante el proceso de lectura.

Los archivos Excel, por su parte, ofrecen capacidades más sofisticadas para el almacenamiento y organización de datos. Su estructura más compleja permite el manejo de múltiples hojas de cálculo, fórmulas y formatos específicos. La lectura de estos archivos requiere una comprensión profunda de su estructura interna y la capacidad de manejar elementos como referencias entre celdas, fórmulas calculadas y formatos condicionales. El proceso de importación debe considerar cuidadosamente estos aspectos para garantizar la integridad de los datos extraídos.

Tabla 1. Características y consideraciones de formatos comunes

Formato	Ventajas	Desventajas	Casos de uso típicos
CSV	Simple y universal, tamaño reducido, fácil de procesar.	Limitado a datos tabulares, problemas con caracteres especiales.	Intercambio de datos simples, logs, datos transaccionales.
Excel (.xlsx)	Soporte para múltiples hojas, fórmulas, formato rico.	Mayor tamaño de archivo puede contener macros no deseados.	Reportes financieros, datos con múltiples categorías.
JSON	Flexible, jerárquico, popular en web.	Puede ser verbose, más complejo de procesar.	APIs web, datos semiestructurados.
XML	Altamente estructurado, autodescriptivo.	Sintaxis compleja, overhead significativo.	Intercambio de datos empresariales, configuraciones.
TXT	Máxima compatibilidad, simple.	Sin estructura inherente, formato limitado.	Logs, datos no estructurados, texto plano.

Fuente. OIT, 2024.

2.2. Conexión y extracción desde bases de datos

La conexión con bases de datos representa uno de los aspectos primarios en la obtención de datos para análisis. Este proceso requiere un enfoque sistemático que abarca desde la configuración inicial hasta la ejecución de consultas complejas, asegurando en todo momento la eficiencia y seguridad en el acceso a los datos.

- **Configuración de conexiones**

  El proceso de configuración de conexiones a bases de datos comienza con la especificación precisa de los parámetros de conexión. Estos incluyen la

dirección del servidor, el puerto de conexión, el nombre de la base de datos y las credenciales de acceso. La construcción de cadenas de conexión debe seguir las mejores prácticas de seguridad, evitando la exposición directa de credenciales en el código y utilizando variables de entorno o sistemas de gestión de secretos para manejar información sensible.

La implementación de pools de conexiones constituye una estrategia fundamental para optimizar el rendimiento y la gestión de recursos. Un pool de conexiones mantiene un conjunto de conexiones preestablecidas que pueden reutilizarse, eliminando la sobrecarga asociada con la creación y destrucción frecuente de conexiones. La configuración adecuada del tamaño del pool es determinante: un pool demasiado pequeño puede crear cuellos de botella, mientras que uno excesivamente grande puede desperdiciar recursos del sistema.

Los aspectos de seguridad en la configuración de conexiones requieren especial atención. La implementación de SSL/TLS para el cifrado de comunicaciones, la configuración de firewalls y la gestión de certificados digitales son elementos esenciales para garantizar una conexión segura. Además, la implementación de políticas de timeout y retry ayuda a manejar situaciones de conectividad intermitente o problemas de red.

- **Consultas básicas y avanzadas**

  El dominio de las consultas SQL, desde las más básicas hasta las más sofisticadas, resulta fundamental para la extracción efectiva de datos. Las consultas básicas constituyen el fundamento de la interacción con bases de datos relacionales. La sentencia SELECT, pilar de las consultas de recuperación, permite especificar exactamente qué datos se desean

obtener. La cláusula WHERE facilita el filtrado preciso de registros, mientras que ORDER BY permite controlar la ordenación de los resultados. Las operaciones de agregación mediante GROUP BY, combinadas con funciones como COUNT, SUM, AVG, MAX y MIN, permiten obtener resúmenes estadísticos valiosos de los datos.

Las consultas avanzadas expanden significativamente las capacidades de extracción y transformación de datos. Los JOINS representan una herramienta fundamental para combinar datos de múltiples tablas. Las uniones pueden ser internas (INNER JOIN), externas (LEFT, RIGHT, FULL OUTER JOIN) o cruzadas (CROSS JOIN), cada una con sus propios casos de uso y consideraciones de rendimiento. La selección del tipo adecuado de JOIN es clave para mantener la integridad de los datos y optimizar el rendimiento de las consultas.

Las subconsultas o consultas anidadas permiten realizar operaciones más complejas, facilitando la implementación de lógica de negocio sofisticada. Estas pueden aparecer en la cláusula SELECT, FROM, o WHERE, y pueden ser correlacionadas o no correlacionadas. El uso efectivo de subconsultas requiere una comprensión profunda de su impacto en el rendimiento y las alternativas disponibles, como las Common Table Expressions (CTEs).

Las funciones de ventana (WINDOW FUNCTIONS) representan una característica avanzada particularmente útil en el análisis de datos. Estas funciones permiten realizar cálculos a través de conjuntos de filas relacionadas con la fila actual, sin necesidad de agrupar los resultados. Funciones como ROW_NUMBER(), RANK(), LAG(), y LEAD() facilitan

análisis sofisticados como el cálculo de diferencias entre filas consecutivas o la identificación de patrones en series temporales.

La optimización de consultas constituye también es altamente relevante en el trabajo con bases de datos. Esto implica la comprensión y utilización efectiva de índices, la minimización de operaciones costosas como los CROSS JOINS, y el uso apropiado de hints cuando sea necesario. El análisis de planes de ejecución es fundamental para identificar y resolver problemas de rendimiento en consultas complejas.

Los procedimientos almacenados y las funciones definidas por el usuario pueden utilizarse para encapsular lógica de consulta compleja y mejorar la reusabilidad del código. Sin embargo, su uso debe equilibrarse con la necesidad de mantener la lógica de negocio accesible y modificable desde las aplicaciones cliente.

2.3. Mejores prácticas en la importación

La implementación de mejores prácticas en la importación de datos busca garantizar la calidad y eficiencia del proceso de análisis. La validación de datos durante la importación constituye una primera línea de defensa contra errores y anomalías que podrían afectar análisis posteriores.

El manejo de errores durante la importación debe ser robusto y contemplar diferentes escenarios problemáticos. Esto incluye la gestión de formatos incorrectos, valores faltantes, problemas de codificación y errores de conexión. La implementación de estrategias de manejo de errores adecuadas permite identificar y resolver problemas tempranamente en el proceso de análisis.

La documentación detallada del proceso de importación resulta medular para la reproducibilidad y mantenimiento del análisis. Esta documentación debe incluir detalles sobre las fuentes de datos, los parámetros de conexión utilizados, las transformaciones aplicadas durante la importación y cualquier consideración especial que deba tenerse en cuenta.

La optimización del rendimiento en la importación de datos requiere considerar factores como el tamaño de los lotes de lectura, la gestión de la memoria y la paralelización de operaciones cuando sea posible. La selección de las estrategias de optimización apropiadas depende tanto de las características de los datos como de las capacidades del sistema.

El control de versiones de los datos importados representa otra consideración importante, especialmente en entornos donde los datos se actualizan regularmente. La implementación de un sistema de control de versiones permite rastrear cambios en los datos y mantener la consistencia en los análisis a lo largo del tiempo.

3. Herramientas y tecnologías para la transformación de datos

La transformación de datos también soporta el proceso de análisis, es en este punto donde los datos crudos se convierten en información procesable y significativa. Este proceso requiere un conjunto diverso de técnicas y metodologías que permiten adaptar, modificar y reestructurar los datos para satisfacer las necesidades específicas del análisis. En este capítulo se exploran las principales técnicas de transformación, desde la conversión básica de tipos de datos hasta operaciones más complejas de clasificación y ordenamiento.

3.1. Conversión de tipos de datos

La conversión de tipos de datos representa una operación fundamental en el proceso de transformación. Esta tarea, aparentemente simple, requiere una comprensión profunda de los diferentes tipos de datos y sus características, así como de las implicaciones y posibles riesgos asociados con cada conversión.

Los tipos de datos numéricos requieren especial atención durante el proceso de conversión. La transformación entre enteros y números de punto flotante, por ejemplo, puede introducir problemas de precisión que deben manejarse cuidadosamente. La pérdida de precisión en cálculos financieros o científicos puede tener consecuencias significativas en los resultados del análisis.

La conversión de fechas y horas presenta desafíos particulares debido a la diversidad de formatos existentes y las consideraciones de zonas horarias. El manejo adecuado de estos tipos de datos requiere atención a detalles como el formato de entrada, la zona horaria de referencia y el manejo de casos especiales como años bisiestos o cambios de horario de verano.

Las conversiones de tipos de texto a numéricos deben implementarse con robustez para manejar casos excepcionales. Esto incluye la gestión de diferentes separadores decimales, formatos regionales y caracteres especiales. Una estrategia común es la implementación de validaciones previas a la conversión para identificar posibles problemas:

- Validación de formato (asegurar que el texto sigue un patrón válido).
- Detección de valores nulos o vacíos.
- Identificación de caracteres no permitidos.
- Verificación de rangos válidos.

Tabla 2. Consideraciones en la conversión de tipos de datos

Tipo de conversión	Consideraciones clave	Riesgos potenciales	Estrategias de mitigación
Texto a número.	Formato regional (p. ej., ".", "," como separador decimal), separadores de miles.	Pérdida de precisión, valores inválidos (p. ej., "1,000.5" con "," como separador decimal).	Validación previa con expresiones regulares, manejo de excepciones, funciones de conversión específicas del lenguaje.
Número a texto.	Precisión decimal, formato de salida (p. ej., número de decimales, separadores de miles).	Truncamiento, pérdida de información significativa.	Especificación explícita del formato de salida, redondeo controlado.
Texto a fecha.	Formatos de entrada (p. ej., "dd/mm/aaaa", "aaaa-mm-dd"), zonas horarias.	Ambigüedad en fechas (p. ej., "01/02/2023"), datos inválidos (p. ej., "30/02/2023").	Parsing robusto con bibliotecas especializadas, validación de formato con expresiones regulares o funciones

Tipo de conversión	Consideraciones clave	Riesgos potenciales	Estrategias de mitigación
			de validación de fechas.
Número a fecha.	Epoch o sistema de referencia temporal, unidades de tiempo (p. ej., segundos, milisegundos).	Interpretación incorrecta de la unidad de tiempo, desbordamiento (valores fuera del rango representable).	Validación de rango, documentación clara del sistema de referencia y unidades.
Categorías a números.	Esquema de codificación (p. ej., one-hot encoding, label encoding), orden de las categorías.	Pérdida de contexto o relaciones entre categorías, introducción de sesgo en el orden.	Mapeo explícito entre categorías y números, preservación de metadatos sobre el esquema de codificación.

Fuente. OIT, 2024.

3.2. Clasificación y ordenamiento

La clasificación y ordenamiento de datos impacta directamente en la calidad y utilidad del análisis posterior. Estas operaciones van más allá del ordenamiento alfabético o numérico, abarcando técnicas sofisticadas que consideran múltiples criterios y relaciones complejas entre dato

El proceso de clasificación implica la organización sistemática de datos en categorías significativas. Este proceso requiere una comprensión profunda del dominio del problema y de las relaciones inherentes entre los datos. La clasificación puede basarse en criterios simples o en combinaciones complejas de múltiples atributos. Los métodos de clasificación pueden variar según la naturaleza de los datos:

a) **Para datos numéricos**

- Clasificación por rangos.
- Categorización por percentiles.
- Agrupación por intervalos.
- Normalización y estandarización.
- Discretización basada en frecuencia.

b) **Para datos categóricos**

- Agrupación por jerarquías.
- Clasificación por frecuencia.
- Consolidación de categorías.
- Mapeo a taxonomías estándar.
- Reducción de cardinalidad.

El ordenamiento de datos, por su parte, requiere la definición clara de criterios de ordenación y la gestión adecuada de casos especiales. La implementación de esquemas de ordenamiento debe considerar aspectos como la estabilidad del ordenamiento, el manejo de valores nulos y la eficiencia computacional en grandes conjuntos de datos.

Las estrategias de ordenamiento múltiple, donde los datos se ordenan según varios criterios en secuencia, requieren una consideración cuidadosa del orden de precedencia y las reglas de desempate. Estas operaciones son particularmente relevantes en análisis que involucran series temporales o datos jerárquicos.

La optimización del rendimiento en operaciones de clasificación y ordenamiento es necesaria cuando se trabaja con grandes volúmenes de datos. Esto puede incluir la

implementación de índices, la selección de algoritmos eficientes y la consideración de estrategias de procesamiento paralelo cuando sea apropiado.

3.3. Eliminación de datos innecesarios

La eliminación de datos innecesarios es una labor de gran cuidado y significancia en el proceso de transformación de datos, la cual contribuye tanto a la calidad del análisis como a la eficiencia del procesamiento. Este proceso requiere un equilibrio o armonía entre la reducción de la redundancia y la preservación de información valiosa.

La identificación de datos redundantes representa el primer paso en este proceso. La redundancia puede manifestarse de diversas formas, desde duplicados exactos hasta redundancias funcionales más sutiles, donde la misma información se representa de diferentes maneras. La detección de estas redundancias requiere un análisis cuidadoso de las relaciones entre variables y la comprensión profunda del contexto del negocio.

La eliminación de variables irrelevantes debe abordarse con particular atención al contexto del análisis. Una variable que parece irrelevante en un contexto puede resultar determinante en otro. El proceso de selección debe considerar factores como la correlación entre variables, la importancia para el negocio y el potencial impacto en análisis futuros. Las decisiones de eliminación deben documentarse cuidadosamente, incluyendo la justificación y las posibles implicaciones.

3.4. Creación de nuevas variables

La creación de nuevas variables representa una oportunidad para enriquecer el análisis mediante la generación de características derivadas que capturen información relevante no presente explícitamente en los datos originales. Este proceso requiere

creatividad y comprensión profunda tanto del dominio del problema como de las técnicas analíticas disponibles.

Las transformaciones matemáticas constituyen una fuente común de nuevas variables. Estas pueden incluir operaciones básicas entre variables existentes, transformaciones no lineales, o cálculos más complejos basados en fórmulas específicas del dominio. La selección de transformaciones apropiadas debe guiarse por consideraciones teóricas y prácticas, incluyendo la interpretabilidad de las nuevas variables y su utilidad para el análisis posterior.

La agregación temporal representa otra técnica valiosa para la creación de variables derivadas. En el análisis de series temporales, por ejemplo, la creación de medias móviles, tasas de cambio o indicadores de estacionalidad puede proporcionar insights valiosos sobre patrones y tendencias en los datos. La selección de ventanas temporales apropiadas y métodos de agregación requiere consideración cuidadosa del contexto y los objetivos del análisis.

3.5. Funciones de transformación esenciales

Las funciones de transformación constituyen el conjunto de herramientas fundamentales que todo analista de datos debe dominar para realizar transformaciones efectivas. Estas funciones abarcan desde operaciones simples hasta transformaciones complejas que pueden aplicarse en diversos contextos analíticos.

Las funciones de normalización y estandarización se emplean para hacer comparables variables con diferentes escalas. La normalización min-max, la estandarización z-score y otras técnicas similares permiten llevar variables a rangos comparables mientras preservan las relaciones relativas entre observaciones. La

selección de la técnica apropiada debe considerar la distribución de los datos y los requisitos específicos del análisis posterior.

Las funciones de manejo de valores atípicos representan otro conjunto clave de herramientas. Estas funciones pueden incluir métodos de winsorización, que limitan valores extremos a ciertos percentiles, o técnicas de transformación robusta que reducen la influencia de valores atípicos en el análisis. La implementación de estas funciones debe equilibrar la necesidad de manejar valores extremos con la importancia de preservar información relevante.

Las funciones de imputación de valores faltantes constituyen una categoría rotunda de transformaciones. Estas pueden variar desde métodos simples como la imputación por media o mediana, hasta técnicas más sofisticadas basadas en modelos predictivos o métodos de imputación múltiple. La selección del método de imputación debe considerar el mecanismo de ausencia de datos, el patrón de valores faltantes y el impacto potencial en el análisis posterior.

Tabla 3. Funciones de transformación esenciales y sus aplicaciones

Tipo de Función	Propósito	Consideraciones clave	Ejemplos de aplicación
Normalización.	Estandarizar escalas de variables a un rango común (p. ej., 0-1).	Preservar las relaciones entre valores, impacto de outliers en la transformación.	Preparación de datos para modelos de Machine Learning, facilitar la comparación de variables con diferentes unidades.
Agregación.	Resumir datos a un nivel superior de granularidad (p. ej., de diario a mensual).	Elección de la función de agregación (p. ej., suma, promedio,	Análisis de series temporales, creación de reportes a nivel gerencial.

Tipo de Función	Propósito	Consideraciones clave	Ejemplos de aplicación
		máximo), pérdida de detalle individual.	
Imputación.	Manejar valores faltantes en un conjunto de datos.	Mecanismo de ausencia de datos (aleatorio, no aleatorio), posible introducción de sesgo.	Asegurar la completitud de los datos para análisis estadístico, aplicación de algoritmos que no toleran valores faltantes.
Discretización.	Convertir una variable continua en categórica.	Número de intervalos o puntos de corte, pérdida de información por la agrupación, interpretabilidad de las categorías resultantes.	Creación de rangos de edad (p. ej., "joven", "adulto", "mayor"), análisis de datos con histogramas.
Codificación.	Transformar variables categóricas en una representación numérica.	Cardinalidad de la variable (número de categorías), problema de "sparse data" con one-hot encoding.	One-hot encoding para variables nominales, label encoding o encoding ordinal para variables ordinales.

Fuente. OIT, 2024.

La aplicación efectiva de estas funciones de transformación requiere tanto comprensión técnica como también juicio analítico para seleccionar las transformaciones más apropiadas en cada contexto. La documentación clara de las transformaciones aplicadas y sus parámetros es decisiva para la reproducibilidad del análisis y la interpretación correcta de los resultados.

La automatización de transformaciones frecuentes mediante funciones personalizadas o `pipelines` de procesamiento puede mejorar significativamente la eficiencia y consistencia del proceso de transformación. Sin embargo, esta automatización debe implementarse con cuidado, incluyendo validaciones apropiadas y mecanismos de manejo de excepciones.

4. Mejores prácticas y casos de aplicación

En el ecosistema actual de la ciencia de datos, la selección y utilización efectiva de herramientas y tecnologías apropiadas impulsan la transformación de datos. Este capítulo explora las diversas soluciones disponibles, desde software especializado hasta plataformas integradas, proporcionando una comprensión importante de sus capacidades y aplicaciones.

4.1. Herramientas especializadas

La transformación de datos requiere herramientas robustas y confiables que puedan manejar eficientemente grandes volúmenes de información mientras mantienen la integridad y calidad de los datos. El mercado actual ofrece una amplia gama de soluciones, desde productos comerciales altamente especializados hasta alternativas de código abierto flexibles y personalizables.

- **Software comercial**. El software comercial para transformación de datos ha evolucionado significativamente en las últimas décadas, ofreciendo soluciones integrales que combinan potencia, facilidad de uso y soporte profesional. Estas herramientas suelen destacar por su interfaz gráfica intuitiva, capacidades avanzadas de procesamiento y robustas características de seguridad y cumplimiento normativo.
- **Alteryx**. Representa uno de los líderes en el espacio de preparación y análisis de datos empresariales. Esta plataforma se distingue por su enfoque en la automatización de flujos de trabajo y su capacidad para manejar datos de múltiples fuentes. Su interfaz visual para el diseño de procesos de transformación permite a los usuarios desarrollar flujos de trabajo complejos sin necesidad de programación extensiva. La

herramienta sobresale en escenarios empresariales donde la velocidad de implementación y la facilidad de uso son prioritarias.

- **Informatica PowerCenter**. Se posiciona como una solución empresarial robusta para la integración y transformación de datos. Su arquitectura escalable y sus capacidades avanzadas de procesamiento la hacen particularmente adecuada para entornos empresariales de gran escala. La plataforma ofrece funcionalidades comprensivas para el mapeo de datos, la gestión de calidad y la documentación de transformaciones, aunque su complejidad puede requerir una inversión significativa en capacitación y recursos especializados.
- **Talend Data Integration**. Aunque ofrece una versión de código abierto, se destaca en su edición comercial por sus capacidades empresariales avanzadas. La plataforma combina una interfaz gráfica intuitiva con la posibilidad de desarrollo de código personalizado, permitiendo a las organizaciones implementar transformaciones complejas mientras mantienen la facilidad de mantenimiento. Sus capacidades de integración con diversas fuentes de datos y sistemas empresariales la hacen particularmente valiosa en entornos heterogéneos.

Soluciones de código abierto.

El ecosistema de código abierto ha producido herramientas poderosas y flexibles para la transformación de datos, ofreciendo alternativas viables a las soluciones comerciales. Estas herramientas se caracterizan por su transparencia, flexibilidad y capacidad de personalización, aunque pueden requerir mayor experiencia técnica para su implementación efectiva.

- **Apache NiFi**

  Se ha establecido como una plataforma robusta para la automatización y gestión de flujos de datos. Su arquitectura distribuida y su modelo de procesamiento basado en flujos lo hacen particularmente adecuado para escenarios que requieren procesamiento en tiempo real y alta disponibilidad. La plataforma ofrece una interfaz web intuitiva para el diseño de flujos de datos, mientras mantiene capacidades avanzadas de monitoreo y control de versiones.

- **OpenRefine**

  Anteriormente conocido como Google Refine, proporciona una herramienta específicamente diseñada para la limpieza y transformación de datos desordenados. Su interfaz similar a una hoja de cálculo, combinada con capacidades avanzadas de clustering y normalización, lo hace especialmente útil para tareas de limpieza de datos y normalización de valores. La herramienta destaca en escenarios que requieren exploración interactiva y refinamiento iterativo de datos.

- **Pentaho Data Integration (PDI)**

  También conocido como Kettle, ofrece una suite completa de herramientas para la integración y transformación de datos. Su interfaz gráfica para el diseño de transformaciones, combinada con un amplio conjunto de componentes predefinidos, facilita el desarrollo de procesos de transformación complejos. La plataforma se destaca por su capacidad para manejar diversos formatos de datos y su extensibilidad mediante plugins personalizados.

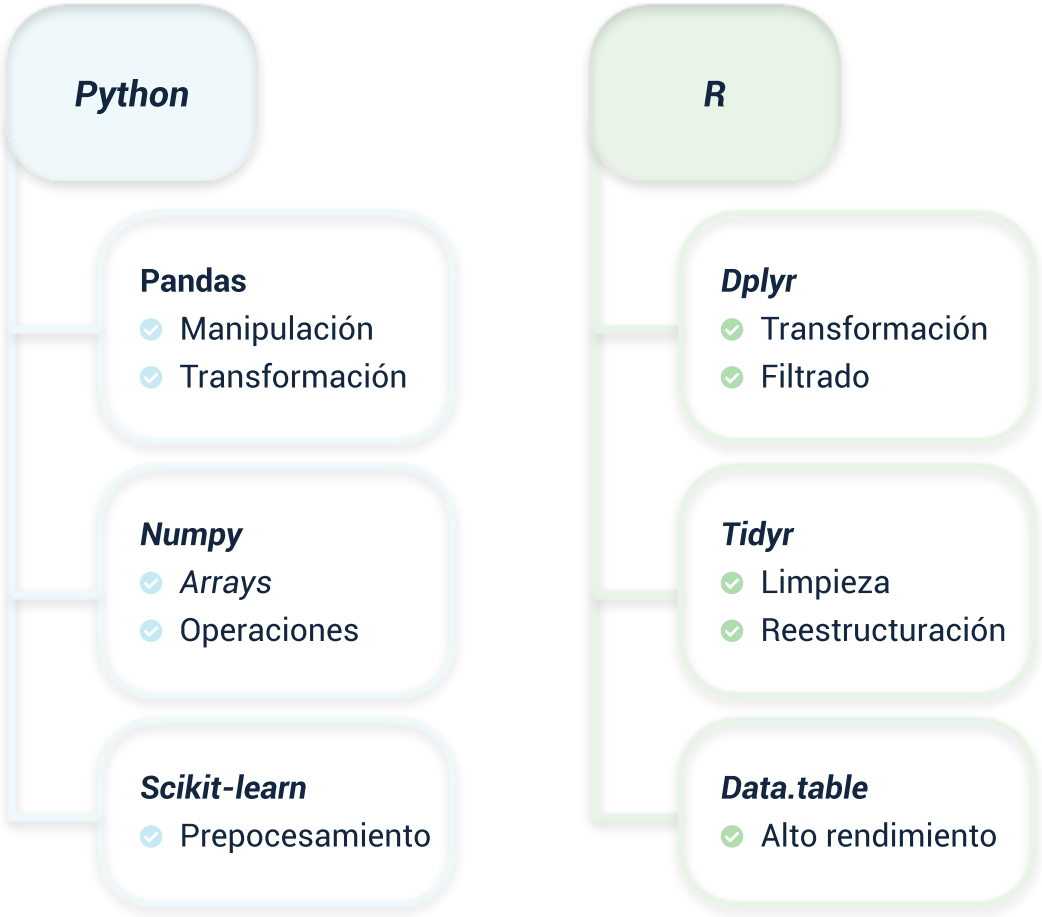
La selección entre soluciones comerciales y de código abierto debe considerar múltiples factores, incluyendo los requisitos específicos del proyecto, las restricciones presupuestarias, la experiencia técnica disponible y las consideraciones de soporte y mantenimiento a largo plazo. Las organizaciones frecuentemente optan por un enfoque híbrido, combinando herramientas comerciales y de código abierto según las necesidades específicas de cada caso de uso.

La evolución continua tanto de las soluciones comerciales como de código abierto refleja la creciente importancia de la transformación de datos en el panorama analítico actual. La tendencia hacia la automatización, la integración de capacidades de machine learning y la mejora en la usabilidad continúa impulsando innovaciones en ambos segmentos del mercado.

4.2. Bibliotecas principales en R y Python

El ecosistema de bibliotecas para transformación de datos en R y Python ha evolucionado significativamente, proporcionando herramientas poderosas y flexibles que facilitan el procesamiento eficiente de datos. Estas bibliotecas constituyen el núcleo de muchas soluciones modernas de transformación de datos, ofreciendo tanto flexibilidad como rendimiento.

Figura 2. Proceso general de la analítica de datos



Fuente. OIT, 2024.

En el entorno Python, pandas se ha establecido como la biblioteca fundamental para la manipulación y transformación de datos. Su estructura principal, el DataFrame, proporciona una interfaz intuitiva y eficiente para operaciones complejas de transformación. La biblioteca ofrece funcionalidades avanzadas para limpieza, reestructuración y agregación de datos, mientras mantiene un rendimiento optimizado para grandes conjuntos de datos.

NumPy complementa a pandas proporcionando capacidades fundamentales para operaciones numéricas y manipulación de arrays. Su integración cercana con pandas permite operaciones vectorizadas eficientes y transformaciones matemáticas complejas.

En el ecosistema R, el conjunto de paquetes tidyverse, particularmente dplyr y tidyr, ha revolucionado la forma en que se realizan las transformaciones de datos. dplyr proporciona una gramática consistente y expresiva para manipulación de datos, mientras que tidyr se especializa en la restructuración y limpieza de datos desordenados.

La biblioteca data.table en R se destaca por su rendimiento excepcional en operaciones con grandes conjuntos de datos. Su sintaxis concisa y eficiente la hace particularmente valiosa en escenarios que requieren procesamiento de alto rendimiento.

4.3. Automatización de procesos de transformación

La automatización de procesos de transformación es una tendencia evidente en la modernización de los flujos de trabajo de datos. Esta automatización no solo mejora la eficiencia operativa, sino que también reduce errores y garantiza la consistencia en el procesamiento de datos.

Los pipelines de transformación automatizados constituyen la columna vertebral de muchos sistemas modernos de procesamiento de datos. Estos pipelines pueden implementarse utilizando diversas tecnologías, desde scripts simples hasta sistemas orquestados complejos. La selección de la estrategia de automatización

depende de factores como la frecuencia de actualización de datos, los requisitos de rendimiento y la complejidad de las transformaciones.

Apache Airflow se ha convertido en una herramienta líder para la orquestación de flujos de trabajo de datos. Su modelo basado en DAG (Grafos Acíclicos Dirigidos) permite definir dependencias complejas entre tareas mientras mantiene la flexibilidad para manejar fallos y reintentos. La plataforma proporciona capacidades robustas de monitoreo y logging, facilitando la identificación y resolución de problemas en los procesos automatizados.

La implementación de pruebas automatizadas en los procesos de transformación es una cuestión primordial para mantener la calidad de los datos. Estas pruebas pueden incluir validaciones de esquema, verificaciones de integridad y pruebas de lógica de negocio. La automatización de estas pruebas ayuda a identificar problemas tempranamente en el proceso de transformación.

La documentación automatizada de los procesos de transformación, incluyendo el linaje de datos y los metadatos de transformación, facilita el mantenimiento y la auditoría de los sistemas de procesamiento de datos. Herramientas como Great Expectations permiten definir y validar automáticamente expectativas sobre los datos, proporcionando documentación viva que evoluciona con los datos.

5. Mejores prácticas y casos de aplicación

La implementación exitosa de procesos de limpieza y transformación de datos requiere el dominio técnico de herramientas y metodologías, al igual que la adhesión a mejores prácticas y la comprensión de su aplicación en contextos reales. Este capítulo explora las prácticas fundamentales y presenta casos de estudio que ilustran su implementación efectiva en situaciones del mundo real.

5.1. Documentación y control de calidad

La documentación exhaustiva y el control de calidad riguroso constituyen pilares fundamentales en cualquier proceso de transformación de datos. Una documentación efectiva no solo facilita el mantenimiento y la reproducibilidad de los procesos, sino que también proporciona transparencia y trazabilidad en las transformaciones realizadas.

La documentación de procesos de transformación debe abordar múltiples niveles de detalle. A nivel estratégico, debe incluir los objetivos y justificación de las transformaciones realizadas, estableciendo claramente el contexto empresarial y los requisitos del negocio. A nivel técnico, debe detallar las especificaciones de las transformaciones, incluyendo:

- Descripción detallada de las fuentes de datos.
- Mapeo de transformaciones y reglas de negocio aplicadas.
- Tratamiento de casos especiales y excepciones.
- Dependencias y requisitos técnicos.
- Procedimientos de validación implementados.

El control de calidad en la transformación de datos debe implementarse como un proceso continuo y sistemático. Los elementos clave de un sistema robusto de control de calidad incluyen:

a) Validaciones técnicas

- Verificación de integridad de datos.
- Comprobación de consistencia en tipos de datos.
- Validación de rangos y restricciones.
- Detección de anomalías y valores atípicos.

b) Validaciones de negocio

- Conformidad con reglas de negocio.
- Verificación de relaciones entre entidades.
- Comprobación de agregaciones y cálculos.
- Validación de tendencias y patrones esperados.

La implementación de pruebas automatizadas constituye una práctica esencial para mantener la calidad de las transformaciones a lo largo del tiempo. Estas pruebas deben ejecutarse de manera regular y documentar sus resultados de forma sistemática, permitiendo la detección temprana de problemas y la verificación continua de la calidad de los datos.

5.2. Casos de estudio prácticos

La aplicación práctica de técnicas de transformación de datos se ilustra mejor a través de casos de estudio reales que demuestran tanto los desafíos enfrentados como las soluciones implementadas.

a) **Caso 1: Integración de datos de ventas minoristas**

Una cadena minorista nacional enfrentaba el desafío de integrar datos de ventas provenientes de múltiples sistemas punto de venta, cada uno con sus propias particularidades en formato y estructura. El proceso de transformación implementado incluía:

Desafíos principales:

- Inconsistencia en formatos de fecha y hora entre sistemas.
- Variaciones en la codificación de productos.
- Discrepancias en las convenciones de nombres.

Solución implementada:

- Desarrollo de un pipeline automatizado de transformación.
- Implementación de mapeos estandarizados para códigos de productos.
- Creación de un sistema de validación multinivel.
- Establecimiento de procesos de reconciliación diaria.

Resultados:

- Reducción del 75% en el tiempo de procesamiento.
- Mejora significativa en la precisión de los reportes.
- Mayor confiabilidad en el análisis de tendencias.

b) **Caso 2: Limpieza de datos de investigación médica**

Un instituto de investigación médica requería la consolidación y limpieza de datos de ensayos clínicos provenientes de múltiples centros de investigación.

Desafíos principales:

- Datos incompletos y valores faltantes.

- Inconsistencias en la codificación de diagnósticos.
- Necesidad de mantener la trazabilidad de todas las transformaciones.

Solución implementada:

- Desarrollo de protocolos estandarizados de limpieza.
- Implementación de sistemas de validación específicos del dominio.
- Creación de registros detallados de transformaciones.
- Establecimiento de procesos de revisión por pares.

Resultados:

- Mejora en la calidad y confiabilidad de los datos.
- Reducción significativa en el tiempo de preparación de datos.
- Mayor facilidad en la reproducción de análisis.

La documentación detallada de estos casos de estudio, incluyendo las lecciones aprendidas y las mejores prácticas identificadas, proporciona insights valiosos para la implementación de soluciones similares en otros contextos. La experiencia acumulada en estos casos demuestra la importancia de:

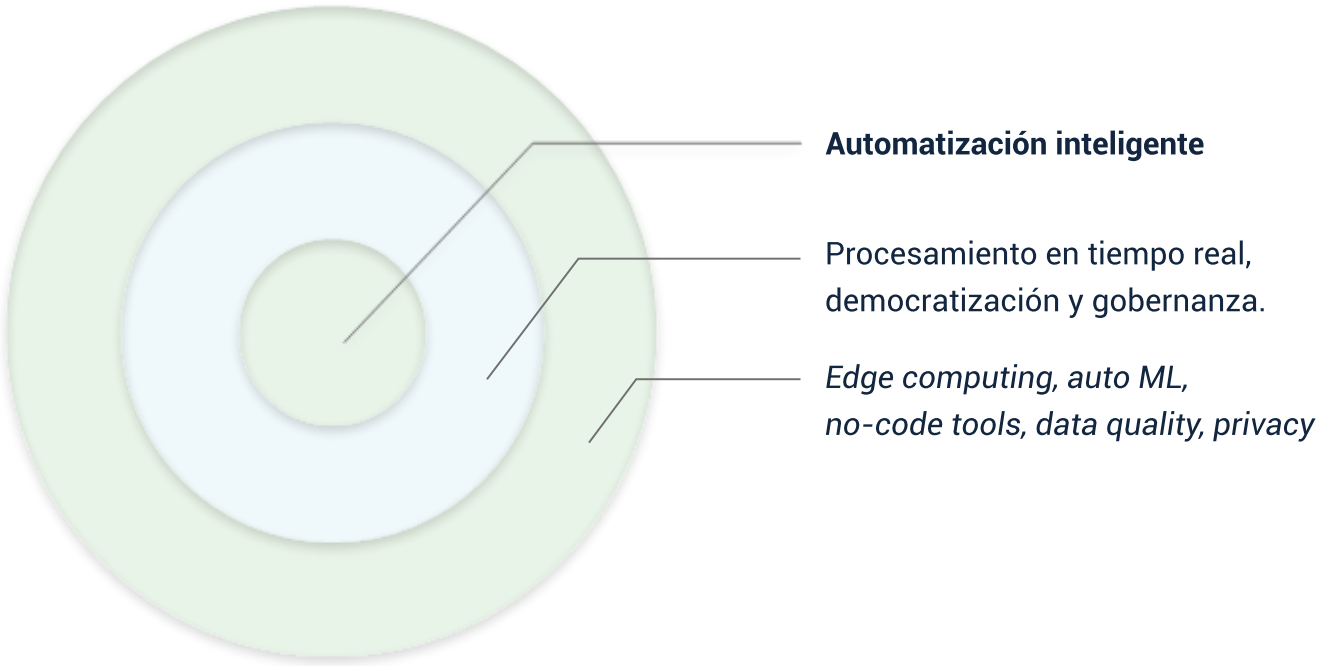
- Establecer procesos claros y documentados.
- Implementar controles de calidad robustos.
- Mantener la trazabilidad de las transformaciones.
- Adaptar las soluciones al contexto específico.
- Fomentar la colaboración entre equipos técnicos y de negocio.

5.3. Tendencias y recomendaciones

El campo de la transformación y limpieza de datos continúa evolucionando rápidamente, impulsado por avances tecnológicos y nuevas necesidades empresariales.

Una de las tendencias más significativas es la creciente adopción de técnicas de automatización basadas en inteligencia artificial para la detección y corrección de anomalías en los datos. Los sistemas de aprendizaje automático están siendo incorporados cada vez más en los procesos de limpieza, permitiendo identificar patrones sutiles y realizar correcciones de manera más eficiente.

Figura 3. Tendencias en transformación de datos



Fuente. OIT, 2024.

La democratización de las herramientas de transformación de datos representa otra tendencia significativa. Las plataformas modernas están evolucionando hacia interfaces más intuitivas que permiten a usuarios con diversos niveles de experiencia técnica participar en los procesos de transformación de datos. Esta tendencia está acompañada por un énfasis creciente en la transparencia y explicabilidad de las transformaciones realizadas.

El procesamiento en tiempo real de datos se está convirtiendo en un requisito cada vez más común. Las organizaciones están migrando hacia arquitecturas que permiten la transformación y limpieza de datos en tiempo real, facilitando la toma de decisiones más ágil y la respuesta inmediata a cambios en los datos.

La gobernanza de datos y el cumplimiento normativo continúan ganando importancia, especialmente en un contexto de regulaciones cada vez más estrictas sobre privacidad y protección de datos. Las organizaciones están implementando frameworks más robustos para asegurar que los procesos de transformación de datos cumplan con requisitos regulatorios mientras mantienen la eficiencia operativa.

5.4. Conclusiones

La limpieza y transformación de datos constituye un elemento fundamental en la cadena de valor de los datos modernos. A medida que las organizaciones continúan su transformación digital, la capacidad para procesar y preparar datos de manera eficiente se vuelve cada vez más crítica para el éxito empresarial.

La adopción de mejores prácticas en documentación y control de calidad, combinada con la implementación de soluciones tecnológicas avanzadas, permite a las organizaciones maximizar el valor de sus datos mientras mantienen altos estándares de calidad y confiabilidad. Los casos de estudio presentados demuestran que el éxito en la transformación de datos requiere un equilibrio entre la excelencia técnica y la comprensión profunda del contexto empresarial.

Las tendencias emergentes sugieren un futuro donde la transformación de datos será cada vez más automatizada e integrada en los procesos empresariales, pero también más accesible y transparente. La continua evolución de herramientas y

metodologías promete facilitar estos procesos, mientras que el énfasis en la gobernanza y la calidad de datos asegura que los resultados sean confiables y utilizables.

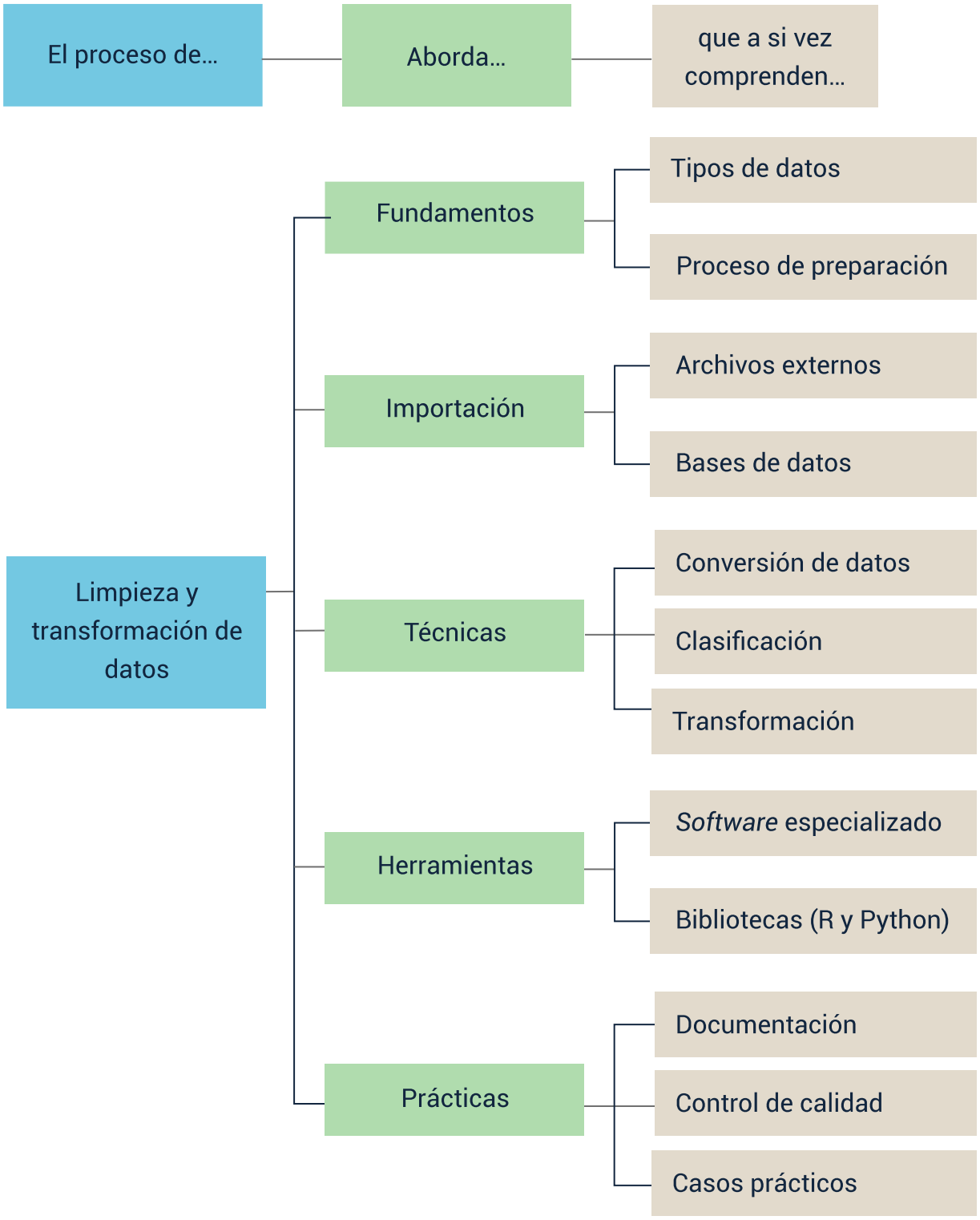
En última instancia, el éxito en la transformación de datos no depende solo de la tecnología utilizada, sino también de la capacidad de las organizaciones para establecer procesos sólidos, mantener altos estándares de calidad y adaptarse a un panorama tecnológico en constante evolución. La inversión en capacitación, herramientas adecuadas y procesos bien diseñados continuará siendo esencial para aprovechar al máximo el potencial de los datos en el entorno empresarial moderno.

Síntesis

El siguiente diagrama proporciona una visión general sintetizada de los principales temas abordados en este componente sobre limpieza y transformación de datos. Este mapa está diseñado para ayudar al lector a visualizar la interconexión entre los diversos elementos que conforman el proceso integral de preparación y transformación de datos.

En el origen del diagrama se encuentra el concepto principal de limpieza y transformación de datos, del cual se ramifican cinco áreas fundamentales: fundamentos, importación de datos, técnicas de transformación, herramientas y mejores prácticas. Cada una de estas áreas se desglosa en subtemas clave, reflejando la estructura y el contenido del componente, desde los conceptos básicos hasta las aplicaciones prácticas y tendencias actuales.

Este diagrama sirve como una guía visual para navegar por los conceptos presentados en el texto, permitiendo al lector comprender rápidamente el flujo y la interrelación de los procesos involucrados en la preparación y transformación de datos. Al revisar este mapa, el aprendiz podrá apreciar cómo los diferentes aspectos, desde la importación inicial hasta las mejores prácticas, se integran para formar un proceso coherente y sistemático. Se invita a explorar este diagrama como un complemento al contenido detallado del componente, utilizándolo como una referencia rápida y un recordatorio visual de los conceptos clave en el campo de la limpieza y transformación de datos.



Fuente. OIT, 2024.

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
1. Fundamentos de la limpieza y transformación de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023b, septiembre 5). Datos sucios.	Video	https://www.youtube.com/watch?v=qf6MR4o58cs
1. Fundamentos de la limpieza y transformación de datos	Limpiar datos de Excel, CSV, PDF y Hojas de cálculo de Google con el intérprete de datos. (s. f.). Tableau.	Portal web	https://help.tableau.com/current/pro/desktop/es-es/data_interpreter.htm
2. Importación y lectura de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023d, septiembre 5). Importación de librerías.	Video	https://www.youtube.com/watch?v=lSvA7RCXkhM
3. Técnicas de transformación de datos.	Ecosistema de Recursos Educativos Digitales SENA. (2023f, septiembre 5). Limpieza y transformación de datos con Python.	Video	https://www.youtube.com/watch?v=jL4cm_0X68Y
4. Herramientas y tecnologías para la transformación de datos.	Ecosistema de Recursos Educativos Digitales SENA. (2023e, septiembre 5). Instalación de Anaconda.	Video	https://www.youtube.com/watch?v=wSdQpgVSPvY
5. Mejores prácticas y casos de aplicación	Ecosistema de Recursos Educativos Digitales SENA. (2023c, septiembre 5). jemplo problemas en la	Video	https://www.youtube.com/watch?v=LOlsg6ZkdcA

Tema	Referencia	Tipo de material	Enlace del recurso
	recolección de la información.		
5. Mejores prácticas y casos de aplicación	Ecosistema de Recursos Educativos Digitales SENA. (2023g, septiembre 5). Puesta en Marcha.	Video	https://www.youtube.com/watch?v=OanYK6mqBIo

Glosario

Análisis exploratorio: proceso inicial de investigación de datos para descubrir patrones, detectar anomalías y verificar suposiciones mediante estadísticas resumidas y representaciones gráficas.

Data Frame: estructura de datos bidimensional que organiza los datos en filas y columnas, similar a una hoja de cálculo o tabla de base de datos. Es fundamental en lenguajes como R y Python para el análisis de datos.

ETL (Extract, Transform, Load): proceso que involucra la extracción de datos de diversas fuentes, su transformación para satisfacer necesidades operativas y la carga en un destino final.

Firewall: sistema de seguridad que monitorea y controla el tráfico de red entrante y saliente basado en reglas de seguridad predeterminadas. Es crucial en la protección de conexiones a bases de datos.

Framework: marco de trabajo que proporciona una estructura estandarizada y mejores prácticas para el desarrollo de soluciones. Incluye herramientas, bibliotecas y convenciones predefinidas.

Logging: proceso de registro y almacenamiento de eventos, errores y actividades del sistema, crucial para el monitoreo, depuración y auditoría de procesos de transformación de datos.

Parsing: proceso de analizar una secuencia de símbolos o texto para determinar su estructura gramatical con respecto a una gramática formal. Fundamental en la lectura y procesamiento de archivos.

Pipeline: secuencia de procesos conectados donde la salida de uno es la entrada del siguiente. En transformación de datos, representa un flujo de trabajo automatizado de procesamiento.

Pool de conexiones: caché de conexiones a bases de datos que se mantienen disponibles para reutilización, mejorando el rendimiento y la gestión de recursos en aplicaciones.

Scripts: programas generalmente simples que automatizan la ejecución de tareas que podrían realizarse paso a paso manualmente.

Timestamp: marca temporal que indica la fecha y hora en que ocurrió un evento particular, fundamental en el seguimiento y ordenamiento de datos temporales.

Validación cruzada: técnica para evaluar la calidad y robustez de un modelo estadístico, dividiendo los datos en subconjuntos para prueba y entrenamiento.

Winsorización: técnica estadística para tratar valores atípicos, donde los valores extremos se reemplazan por valores menos extremos en un determinado percentil.

Workflow: flujo de trabajo que representa la secuencia de procesos o pasos necesarios para completar una tarea específica en el procesamiento de datos.

Referencias bibliográficas

Bansal, S. K. & Kagemann, S. (2015). Integrating Big Data: A Semantic Extract-Transform-Load Framework. Computer, 48(3), 42-50. https://doi.org/10.1109/mc.2015.76

De Vries, A. & Meys, J. (2012). R For Dummies. John Wiley & Sons.

Díaz, C. O., Soler, P., Pérez, M. & Mier, A. (2024). OMASHU: La ciencia detrás del éxito; Big Data e IA en los eSports. Revista SISTEMAS, 170, 61-79. https://doi.org/10.29236/sistemas.n170a7

Guardelli, E. (2024). Minería de Procesos: Convertir Datos en Valor. MedTechBiz.

Jones, H. (2018). Analítica de Datos: Una guía esencial para principiantes en minería de datos, recolección de datos, análisis de Big Data para negocios y conceptos de inteligencia empresarial. Independently Published.

Leyva, D. S. (2024). Domina Machine Learning: Guía completa para principiantes. Independently Published.

McKinsey, W. (2023). Python para análisis de datos. Anaya Multimedia.

Orlandi, M. A. M. (2024). Tecnologías Big Data, Minería de Datos y Analítica aplicada a la gestión de Recursos Humanos: contiene: un caso de estudio. Editora Dialética.

Shovic, J. C. & Simpson, A. (2019). Python All-in-One For Dummies. John Wiley & Sons.

Subirats Maté, L., Pérez Trenard, D. O., Calvo González, M. & Isabel Guitart Hormigo. (2019). Introducción a la limpieza y análisis de los datos. https://openaccess.uoc.edu/bitstream/10609/148647/1/IntroduccionALaLimpiezaYAnalisisDeLosDatos.pdf

Créditos

Elaborado por:

Organización Internacional del Trabajo