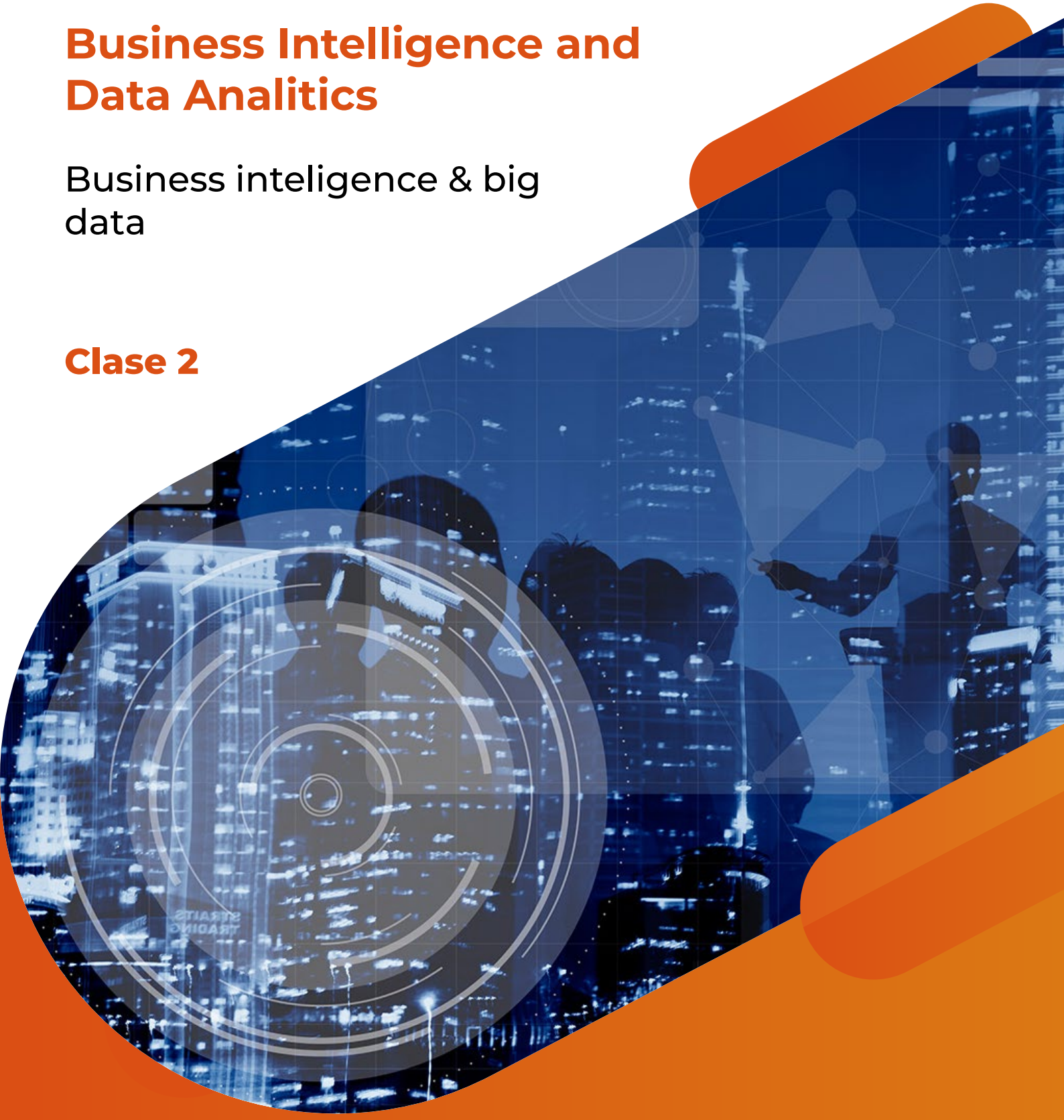
Business Intelligence and Data Analitics

Business inteligence & big data

Clase 2



MAESTRÍA EN ADMINISTRACIÓN DE EMPRESAS MBA

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

La depuración de bases de datos es una etapa crítica en el procesamiento de datos que garantiza la calidad y precisión de la información. Power Query, una herramienta de Microsoft, se ha consolidado como una opción poderosa para la extracción, transformación y carga de datos (ETL, por sus siglas en inglés). Este proceso permite a los usuarios extraer datos de múltiples fuentes, transformarlos mediante operaciones de limpieza y estandarización, y, finalmente, cargarlos en un formato adecuado para el análisis. Según Hernando (2020), las herramientas ETL como Power Query permiten una manipulación de datos eficiente y reproducible, lo que facilita la generación de reportes confiables.

El enfoque ETL dentro de Power Query se centra en la transformación de datos en varios niveles. Durante la fase de extracción, se recuperan datos de distintas fuentes, que pueden incluir bases de datos SQL, archivos Excel o servicios en la nube. En la fase de transformación, Power Query permite a los usuarios aplicar una serie de operaciones como la eliminación de duplicados, el manejo de valores nulos y la conversión de tipos de datos. Estas operaciones son esenciales para garantizar la coherencia de los datos antes de cargarlos en la fase final, donde se almacenan en tablas optimizadas para el análisis (Jones, 2021). La automatización de este proceso asegura la minimización de errores humanos y aumenta la eficiencia en proyectos de análisis de grandes volúmenes de datos.

Clase 2: Depuración de bases de datos con Power Query fundamentada en ETL

Resultado o resultados de aprendizaje que será abordado con el contenido de la clase: Integrar bases de datos estructuradas para transformar datos en información

2. Depuración de bases de datos con Power Query & ETL

La depuración de bases de datos es un paso esencial en la gestión de datos, ya que garantiza la calidad, coherencia y precisión de la información. Power Query, una herramienta desarrollada por Microsoft, facilita la depuración de datos mediante la integración de procesos ETL (Extracción, Transformación y Carga). En el contexto de ETL, Power Query permite la extracción de datos de múltiples fuentes, su transformación para adecuarlos a las necesidades del análisis y la carga final en un formato listo para su uso. Según Torres (2021), la funcionalidad de Power Query simplifica la depuración de datos al automatizar la limpieza de grandes volúmenes de información y mejorar la eficiencia del análisis.

El enfoque de Power Query en los procesos ETL destaca por su capacidad para transformar los datos. Durante la fase de transformación, los usuarios pueden aplicar diversas operaciones de limpieza, como la eliminación de duplicados, la corrección de inconsistencias y el manejo de valores nulos. Estas operaciones no solo optimizan la calidad de los datos, sino que también los preparan para su uso en modelos de análisis y visualización. Como señalan Ruiz y García (2020), la transformación de datos en Power Query permite un control exhaustivo sobre los conjuntos de datos, resultando en un proceso de depuración preciso y estructurado, alineado con las mejores prácticas de ETL.

Además, Power Query facilita la automatización de flujos de trabajo dentro del marco ETL, lo cual es clave en la depuración eficiente de bases de datos. Su capacidad de integración con otras herramientas de Microsoft, como Excel y Power BI, lo convierte en una solución robusta para empresas y profesionales que necesitan manejar grandes volúmenes de datos de manera eficiente. Esta automatización permite realizar tareas repetitivas de manera consistente y reduce errores humanos. Como destaca López (2022), Power Query ha revolucionado los procesos ETL al proporcionar un entorno intuitivo y potente para la depuración y preparación de datos.

En la primera clase, vimos el diseño de una base a partir de nuestra creación de la tabla, pero muchas

veces tenemos que trabajar con bases que no son de nuestra autoría y debemos depurarlas. Es ahí donde se recomienda seguir la metodología ETL, que a continuación se detalla en cada etapa.

2.1 Metología ETL

La metodología ETL (Extract, Transform, Load) es un proceso clave en la integración y gestión de datos, utilizado principalmente en la creación de almacenes de datos (data warehouses) y sistemas de análisis empresarial. Consiste en tres fases: la **extracción** de datos desde diversas fuentes, como bases de datos, archivos o aplicaciones externas; la **transformación** de estos datos, que implica limpiar, depurar y modificar el formato de los datos para que sean consistentes y adecuados para el análisis; y, finalmente, la **carga** de los datos transformados en un sistema de almacenamiento o base de datos destino. Este proceso es esencial para consolidar grandes volúmenes de datos y permitir su análisis eficiente y preciso. La metodología ETL garantiza que los datos sean de alta calidad y estén listos para la toma de decisiones en las organizaciones (Kimball & Ross, 2013).

A continuación, detallamos el procedimiento de cada etapa de la metodología:

Fase del ETL	Descripción	Ejemplo
Extracción (Extract)	Obtención de datos desde diversas fuentes, como bases de datos, archivos CSV, APIs o aplicaciones.	Extraer datos de una base de datos ERP.
Transformación (Transform)	Modificación y limpieza de datos para que sean consistentes y adecuados para el análisis. Incluye normalización, eliminación de duplicados, y conversión de formatos.	Limpiar datos eliminando duplicados y unificando formatos de fecha.
Carga (Load)	Almacenamiento de los datos transformados en un sistema de destino, como un data warehouse o base de datos.	Cargar los datos procesados en un data warehouse para análisis.

***Extract* (Extracción)**

La extracción es la primera fase del proceso ETL (Extracción, Transformación y Carga) y consiste en obtener datos de diversas fuentes para su posterior procesamiento. En esta etapa, los datos pueden provenir de bases de datos estructuradas, archivos planos, sistemas ERP o incluso servicios web. Según López y García (2020), la correcta ejecución de la extracción es crucial para garantizar que toda la información relevante se capture de manera completa y precisa, lo que reduce el riesgo de pérdida de datos valiosos durante el proceso. Además, esta fase implica la identificación de los formatos y estructuras de origen, lo que permite que los datos extraídos estén listos para su transformación y adaptación a un formato adecuado para análisis o almacenamiento en un data *warehouse*.

Para más información sobre la metodología ETL, consulta el siguiente enlace:

https://www.youtube.com/watch?v=gZZrV8iyU7w

Obtener los datos a partir de diferentes fuentes de origen, entre las fuentes más comunes tenemos:

- **Fuentes Comunes**:
  - Bases de datos (SQL, NoSQL)
  - Archivos (CSV, Excel, JSON)

- APIs
- Sistemas ERP, CRM, etc.

➢ **Consideraciones**:

- La extracción puede ser completa (todas las filas) o incremental (solo nuevos datos desde la última extracción).
- Debe manejarse la calidad de los datos desde el inicio.

b. *Transform* (Transformación)

La transformación en el proceso ETL (Extracción, Transformación y Carga) es una etapa crucial en la que los datos extraídos de diversas fuentes son modificados y ajustados para cumplir con los requisitos de análisis y almacenamiento. Durante esta fase, se aplican operaciones como la limpieza de datos, conversión de formatos, agregación y creación de nuevos campos. Según Hernández y Soto (2020), la transformación permite que los datos sean homogéneos y consistentes, facilitando su posterior análisis y asegurando que se ajusten a los estándares de calidad. Esta fase es fundamental para resolver problemas de calidad, como la eliminación de duplicados o el tratamiento de valores faltantes, mejorando así la precisión de los resultados del análisis.

Ya con las bases disponibles, pasamos a la transformación, donde se recomienda determinar:

- **Objetivo**: Limpiar y preparar los datos para el análisis.
- **Tareas Comunes**:
  - **Limpieza de datos**: Eliminación de duplicados, tratamiento de valores nulos, corrección de formatos.
  - **Enriquecimiento**: Agregar datos de otras fuentes o realizar cálculos.
  - **Normalización**: Establecer formatos consistentes (por ejemplo, convertir todas las fechas a un formato estándar).
  - **Agregación**: Resumir datos, como calcular promedios o totales.
  - **Filtrado**: Seleccionar solo las columnas y filas relevantes para el análisis.

a. Load (Carga)

Finalmente procedemos a cargar la base de datos

- **Objetivo**: Cargar los datos transformados en el sistema de destino.
- **Destinos Comunes**:
  - Data Warehouses (almacenes de datos)
  - Bases de datos
  - Sistemas de análisis (como Power BI)

- **Tipos de Carga**:
    - **Carga completa**: Sustituye todos los datos en el destino.
    - **Carga incremental**: Agrega solo los nuevos datos o actualiza los existentes.
- **Consideraciones**:
    - Es crucial monitorear el rendimiento y la calidad de la carga.
    - Asegúrate de que las relaciones y estructuras de datos se mantengan adecuadas

Esta metodología ETL se recomienda realizar en Power Query de Power BI. Con las tablas depuradas, procedemos a realizar la conexión entre tablas en Power Pivot

2.1.1 Conexión de bases de datos estructuradas

Power BI es una herramienta de inteligencia empresarial que permite conectar y visualizar datos provenientes de diversas fuentes, incluidas bases de datos estructuradas. Estas bases de datos, como SQL Server, Oracle o MySQL, contienen datos organizados en tablas con relaciones predefinidas, lo que facilita su integración en Power BI. Según López y Ramírez (2021), la conexión de bases de datos estructuradas a Power BI es un proceso directo gracias a su amplia compatibilidad con múltiples sistemas de gestión de bases de datos (DBMS). Esta conexión permite a los usuarios acceder a grandes volúmenes de información y generar reportes dinámicos en tiempo real, optimizando la toma de decisiones basada en datos precisos.

Una de las principales ventajas de conectar bases de datos estructuradas en Power BI es la capacidad de actualizar automáticamente los datos, lo que garantiza que los informes reflejen siempre la información más reciente. Además, Power BI permite transformar y depurar los datos directamente desde la fuente, facilitando su limpieza antes de utilizarlos en análisis o visualizaciones. Como mencionan González y Pérez (2020), la integración de bases de datos estructuradas en Power BI no solo mejora la eficiencia del análisis de datos, sino que también simplifica el trabajo colaborativo entre equipos de TI y analistas de negocio, al proporcionar una única fuente de verdad para la información empresarial.

a. Iniciar Power BI Desktop en tu computadora.

***b.* Seleccionar la Fuente de Datos**

- En la pestaña de **Inicio**, haz clic en **Obtener datos**. Se abrirá una ventana con diferentes opciones de fuentes de datos.

***c.* Elegir el Tipo de Conexión**

- Selecciona el tipo de fuente de datos que deseas conectar. Algunas opciones comunes son:
    - **Archivo**: Excel, CSV, JSON, etc.
    - **Base de datos**: SQL Server, MySQL, PostgreSQL, etc.

- **Servicios en línea**: SharePoint, Google Analytics, etc.

- Haz clic en **Conectar**.

a. Proporcionar Detalles de Conexión

- Dependiendo de la fuente seleccionada, deberás ingresar detalles específicos:
  - **Archivo**: Selecciona el archivo de tu sistema.
  - **Base de datos**: Proporciona el nombre del servidor, base de datos, y las credenciales (usuario y contraseña).
  - **Servicios en línea**: Sigue las instrucciones para autenticarte.

e. **Navegar por los Datos**

- Una vez establecida la conexión, se abrirá el **navegador** de Power BI, donde podrás ver las tablas y vistas disponibles.
- Selecciona las tablas o datos que deseas importar y haz clic en **Cargar**. Si quieres transformarlos antes, puedes hacer clic en **Transformar datos** para abrir el Editor de Power Query.

2.1.2 Validación de las variables numéricas

En Power Query te permitirá extraer, transformar y cargar datos (ETL, por sus siglas en inglés) de diversas fuentes como archivos Excel, bases de datos, servicios web y más. Su principal ventaja es que permite a los usuarios realizar transformaciones complejas de datos sin necesidad de escribir código, lo que simplifica el trabajo con grandes volúmenes de información. Este proceso de transformación de datos en Power Query implica limpiar, estructurar y preparar los datos para análisis, facilitando la integración de diferentes fuentes y formatos.

Entre los pasos para transformar datos en Power Query, te recomiendo seguir los siguientes puntos:

1. Importar datos: El primer paso es cargar los datos desde una fuente. Power Query ofrece múltiples conectores que permiten extraer datos de fuentes como Excel, SQL Server o incluso desde la web.

2. Explorar y comprender los datos: Después de importar los datos, se puede visualizar una vista preliminar en el "Editor de Power Query", donde el usuario puede revisar la estructura de los datos, los tipos de columnas y detectar posibles inconsistencias o valores faltantes.

3. Aplicar transformaciones:

- Eliminar columnas o filas innecesarias: Es posible reducir los datos eliminando columnas o filas que no aportan valor al análisis.

- Filtrar registros: Los datos se pueden filtrar para incluir solo la información relevante según criterios específicos.

- Cambiar el tipo de datos: Es fundamental asignar los tipos correctos de datos a cada columna (por ejemplo, texto, fecha, número) para asegurar que las operaciones posteriores funcionen correctamente.

- Dividir o combinar columnas: Las columnas se pueden dividir (por ejemplo, separar nombre y apellido) o

combinar (unir varias columnas en una).

- Pivotar o despivotar: Permite reestructurar los datos para que se ajusten al formato necesario para el análisis.

4. Cargar los datos transformados: Una vez que los datos han sido limpiados y transformados, se pueden cargar de nuevo a Excel o Power BI para su análisis.

2.1.3 Validación de las variables categóricas

Power Query es una herramienta poderosa que automatiza la manipulación de datos categóricos, mejorando la eficiencia y asegurando que los conjuntos de datos sean precisos y estén listos para su uso en informes y análisis.

- En el Editor de Power Query, puedes realizar diversas transformaciones, como:
    - Filtrar filas.
    - Cambiar tipos de datos.
    - Combinar tablas.
    - Eliminar columnas innecesarias.

Después de realizar las transformaciones deseadas, haz clic en **Cerrar y aplicar** para volver a Power BI.

En **Power Query** de Power BI, la depuración de tablas es esencial para garantizar la calidad y precisión de los datos antes de utilizarlos en informes. Este proceso permite transformar y limpiar los datos para adaptarlos a las necesidades específicas del análisis. Uno de los primeros pasos es eliminar las columnas y filas innecesarias, lo que reduce el tamaño de los datos y mejora el rendimiento del informe (González, 2021). Además, puedes cambiar los tipos de datos de las columnas para asegurarte de que los valores se interpreten correctamente, ya que Power Query permite convertir datos en formatos como texto, números o fechas (López, 2020).

Otra técnica clave es el manejo de valores faltantes o erróneos, donde Power Query facilita la sustitución de valores nulos o la eliminación de filas incompletas (Carmona, 2022). También es posible dividir columnas, fusionar datos o aplicar filtros avanzados para aislar únicamente los registros relevantes para el análisis. Estos pasos aseguran que los datos sean precisos y coherentes.

Por último, la depuración en Power Query permite realizar un seguimiento de cada transformación mediante el uso del **panel de pasos aplicados**, lo que te otorga la flexibilidad de ajustar o eliminar transformaciones en cualquier punto del proceso sin perder los cambios previos (Martínez, 2021).

Para más información relacionada, consulta el siguiente enlace:

https://www.youtube.com/watch?v=MqAagqzH3YI

2.2 Depuración de bases con Power Query (clave para la calidad de datos)

La depuración de bases de datos es esencial para asegurar la integridad y precisión de la información utilizada en el análisis y la toma de decisiones. Power Query, una herramienta de Microsoft, destaca por su capacidad de simplificar y automatizar el proceso de limpieza y transformación de datos. Esta herramienta permite a los usuarios eliminar duplicados, manejar valores nulos y estandarizar formatos de manera eficiente. Según Pérez y Gómez (2021), Power Query es ampliamente utilizado

en entornos empresariales y académicos debido a su facilidad de uso y su integración con otras plataformas como Excel y Power BI.

Un aspecto fundamental de Power Query es su enfoque en la transformación de datos durante la depuración. La transformación permite que los datos extraídos de diferentes fuentes se adecuen a los requisitos del análisis, garantizando cohesión y homogeneidad en los conjuntos de datos. Por ejemplo, los usuarios pueden unificar diferentes formatos de fechas, eliminar errores comunes como espacios en blanco innecesarios y convertir datos a los tipos adecuados (González, 2020). Estas capacidades permiten a las organizaciones mejorar la calidad de los datos de manera rápida y efectiva, optimizando sus procesos de análisis.

Finalmente, la capacidad de Power Query para automatizar procesos repetitivos de depuración es uno de sus mayores beneficios. Los usuarios pueden crear flujos de trabajo que se actualizan automáticamente cuando cambian las fuentes de datos, ahorrando tiempo y reduciendo el riesgo de errores manuales. Como destacan Ramírez y Salinas (2022), la automatización en Power Query es crucial para la gestión eficiente de grandes volúmenes de datos, convirtiéndolo en una herramienta indispensable para profesionales del análisis de datos y gestores de información.

Para realizar la depuración de la base de datos, se requiere utilizar la interfaz de Power Query; entre las fases más comunes de depuración se encuentra la creación de variables:

- Una vez que estés en el Editor de Power Query, verás la tabla de datos cargada.

1.1.1 **Creación de nuevas columnas**

Para crear la columna debes seguir el siguiente procedimiento:

- Ve a la pestaña **Agregar columna** y selecciona **Columna personalizada**.
- En la ventana que aparece, puedes ingresar una fórmula utilizando M (el lenguaje de Power Query).
- Por ejemplo, para crear una columna que sume dos columnas existentes (Columna1 y Columna2), escribe:

  m

  Copiar código

  [Columna1] + [Columna2]

- Dale un nombre a la nueva columna y haz clic en **Aceptar**.

2.2.2 Uso de columnas Índice

La creación de una columna de índice es una práctica común y útil en el manejo de bases de datos, especialmente cuando se trabaja con grandes volúmenes de información. Este índice actúa como un identificador único para cada fila de datos, lo que facilita la localización y organización de la información. Según García y López (2020), la adición de un índice en un conjunto de datos es esencial para realizar operaciones como el filtrado, ordenación y combinación de tablas. En herramientas como Power Query, la creación de una columna de índice es sencilla y permite establecer un control más preciso sobre los datos, mejorando la eficiencia en su manipulación y análisis.

Además de facilitar la organización de datos, una columna de índice puede ser útil para identificar inconsistencias o duplicados. Al asignar un número único a cada fila, es más fácil detectar errores en la base de datos y aplicar correcciones. Como señalan Pérez y Martín (2021), el uso de una columna de índice en análisis de datos es una práctica recomendada en la limpieza de datos, ya que ayuda a garantizar que no existan duplicidades y que cada registro sea único y rastreable. De esta manera, la creación de una columna de índice se convierte en un paso clave en la depuración y preparación de datos para su posterior análisis.

- Para agregar un índice a tus filas, ve a **Agregar columna** y selecciona **Columna de índice**.
- Esto agregará una nueva columna con números secuenciales.

2.2.3 Consolidación de tablas

En Power BI, puedes consolidad tablas a partir de crear una columna personalizada es un proceso que permite agregar nueva información a partir de datos existentes sin modificar las fuentes originales. Esto se realiza a través del editor de Power Query o directamente en el entorno de creación de informes. Las columnas calculadas pueden utilizarse para realizar operaciones aritméticas, concatenar texto, o extraer información específica de otras columnas. Al escribir fórmulas en DAX (Data Analysis Expressions), los usuarios pueden obtener resultados dinámicos que se ajustan al modelo de datos, facilitando análisis más complejos (Russo & Ferrari, 2020).

Las columnas creadas en Power Query permiten aplicar transformaciones avanzadas a los datos antes de cargarlos en el modelo de Power BI. Esto es particularmente útil cuando se quiere limpiar o preparar los datos para análisis, ya que Power Query tiene herramientas como “Agregar columna de ejemplo” que facilita la creación de nuevas columnas sin necesidad de conocimientos avanzados en fórmulas. Por ejemplo, es posible extraer partes específicas de un texto, transformar formatos o realizar cálculos a partir de columnas existentes, mejorando la flexibilidad en la preparación de los datos (Chapman, 2021).

- En la pestaña **Agregar columna**, selecciona **Columna condicional**.
- Configura las condiciones que deseas aplicar. Por ejemplo, si quieres crear una columna que clasifique edades:
    - Si la Edad es menor de 18, entonces “Menor de edad”.
    - Si la Edad es 18 o más, entonces “Adulto”.
- Define las condiciones y asigna un nombre a la columna.

Referencias citadas en la Clase 2.

Webb, C. (2014). *Power query for power BI and Excel*. Apress.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3ra ed.). Wiley.

Collie, R., & Singh, A. (2015). *Power Pivot and Power BI: The Excel User's Guide to DAX, Power Query, Power BI & Power Pivot in Excel 2010-2016*. Tickling Keys, Inc..

Chapman, M. (2021). Beginning Power BI: A practical guide to self-service data analytics with Excel 2019 and Power BI Desktop. Apress.

López, A. (2022). *Automatización de procesos ETL con Power Query*. Editorial Técnica.

Ruiz, J., & García, M. (2020). *Limpieza y transformación de datos en el entorno empresarial*. Data-Works.

Torres, P. (2021). *Depuración de datos y análisis avanzado con herramientas ETL*. Ediciones Data-Science.

Russo, M., & Ferrari, A. (2020). The Definitive Guide to DAX: Business intelligence with Microsoft Excel, SQL Server Analysis Services, and Power BI. Microsoft Press.

González, L. (2020). Optimización de datos con Power Query: Principios y prácticas. Ediciones Digitales.

Pérez, R., & Gómez, M. (2021). Transformación de datos en entornos empresariales. DataWork Press.

Ramírez, F., & Salinas, J. (2022). Automatización y calidad de datos con Power Query. Innovación y Tecnología.

López, A. (2020). *Transformación de datos con Power Query*. Editorial TechData.

Martínez, L. (2021). *Herramientas avanzadas de Power BI*. Editorial Business Solutions.

Definición de los términos citados en la Clase 2.

Power BI:

Herramienta de análisis de datos desarrollada por Microsoft que permite a los usuarios conectar, transformar y visualizar datos mediante informes interactivos y paneles (dashboards). Es ampliamente utilizada para crear visualizaciones avanzadas y compartir insights empresariales de manera fácil y accesible.

Power Query:

Funcionalidad de Power BI y Excel que permite importar, limpiar, transformar y cargar (ETL) datos desde múltiples fuentes. Proporciona una interfaz gráfica intuitiva para realizar operaciones de transformación de datos sin necesidad de conocimientos avanzados de programación.

ETL (Extract, Transform, Load):

Proceso de manejo de datos que consta de tres etapas: Extracción de datos de diversas fuentes

Profundización Clase 2.

Este video es ideal para identificar la interfaz de Power Query.
CLASE_2_PARTE_1.mp4

Aquí encontrarás las opciones más frecuentes del manejo de columnas.
CLASE_2_PARTE_2.mp4

PUCE Virtual

La excelencia no se improvisa

síguenos