Desarrollo de contenido

Unidad 3

Bases de Datos II

Ingeniería de Software y Datos

IU Digital de Antioquia
INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA

Unidad 3. Diseño de procesos ETL

Introducción Unidad 3

En el mundo de las Tecnología de la Información y la Comunicación, la gestión eficiente y efectiva de datos se ha convertido en un pilar fundamental para el éxito empresarial. La capacidad de extraer, transformar y cargar (ETL) datos de manera estratégica hace parte del diseño de procesos. La integración de datos es esencial para tejer la compleja red de información que impulsa las organizaciones modernas.

En esta tercera y última unidad del curso Bases de Datos II nos centraremos en el diseño de procesos ETL, donde la habilidad para orquestar la transformación de datos se convierte en un catalizador en la toma de decisiones informada. También exploraremos las intrincadas sendas de la Integración de Datos ya que, revelando cómo unificar diversas fuentes de información, se puede potenciar la visión holística de una empresa.

Asimismo, estudiaremos las técnicas que enriquecen el proceso, desentrañando métodos innovadores para optimizar la gestión de datos y asegurar la coherencia de la información.

En este curso, entonces, conocerás más sobre la gestión de datos, donde el diseño de procesos, la integración y las técnicas se entrelazan para forjar un cimiento robusto en la era digital. Esperamos que esta sea una exploración apasionante que cambie la manera en que percibes la gestión de datos en el siglo XXI.

Resultados de aprendizaje de la *Unidad 3*

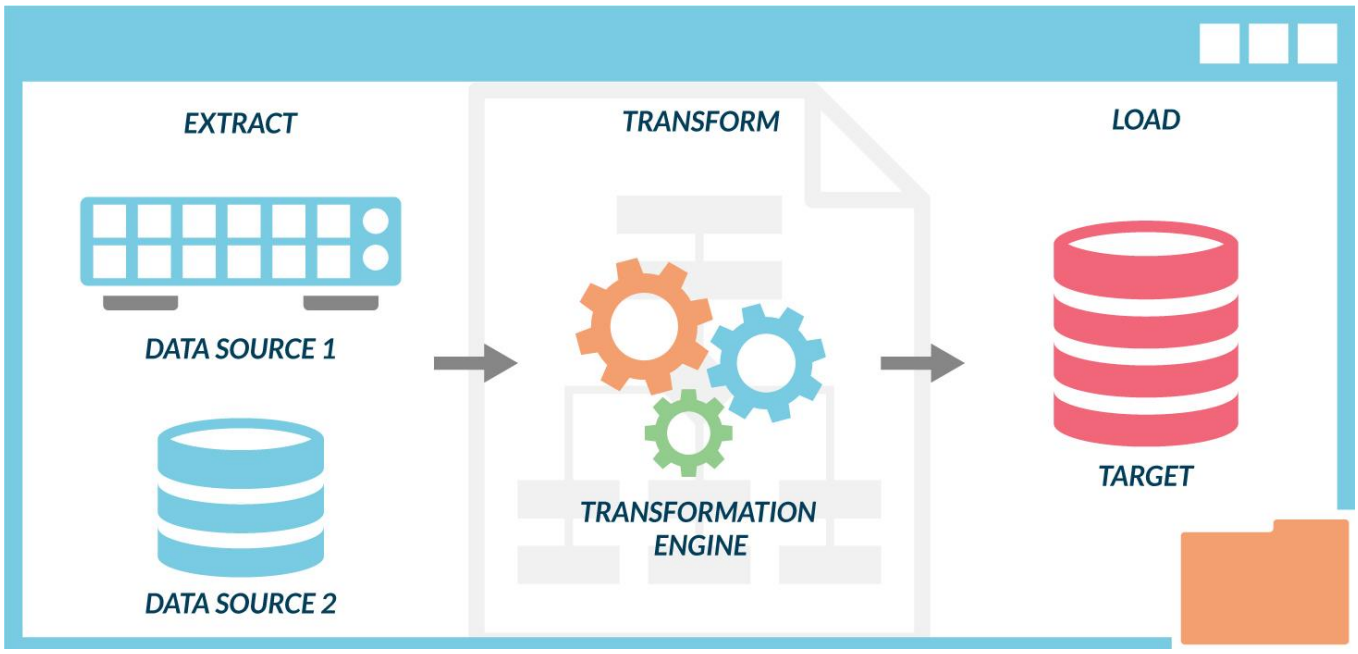
- Analiza los principios básicos de extracción, transformación y carga de datos, así como sus interacciones dentro de un flujo de trabajo ETL.
- Aplica correctamente las técnicas de limpieza, transformación y enriquecimiento de datos para garantizar la coherencia y la calidad de los datos durante el proceso de integración.
- Diseña estrategias de integración de datos que aborden los desafíos específicos de la heterogeneidad de datos y las restricciones de tiempo en un entorno empresarial.

Tema 1. Introducción al diseño de procesos ETL

Es común que las organizaciones se enfrenten al desafío de recolectar datos de diferentes fuentes y en varios formatos, para luego trasladarlos a uno o más almacenes de datos que a veces pueden tener un formato distinto al de origen. En muchos casos, es necesario dar forma o limpiar los datos antes de cargarlos a su destino final.

Sin embargo y a lo largo del tiempo, se han desarrollado diversas herramientas, servicios y procesos para ayudar a superar los desafíos mencionados. Independientemente del método utilizado existe una necesidad general de coordinar el trabajo y aplicar cierto nivel de transformación de datos en la canalización de los mismos (Microsoft.com, 2024).

Figura 1. Proceso ETL



Fuente: adaptado de Microsoft, (2024).

Proceso de extracción, transformación y carga (ETL)

La Extracción, Transformación y Carga (ETL) es una canalización de datos utilizada para reunir información de diversas fuentes. Posteriormente, transforma esos datos de acuerdo con las reglas de negocio y los carga en un almacén de datos de destino. La transformación en el proceso ETL ocurre en un motor especializado y, con frecuencia, implica el uso de tablas de almacenamiento provisional para retener temporalmente los datos, mientras se lleva a cabo la transformación. Finalmente, se cargan en su destino final.

La transformación de datos a menudo conlleva varias operaciones como filtrado, ordenación, agregación, combinación de datos, limpieza de datos, de duplicación y validación de datos.

Frecuentemente, las tres fases del proceso ETL se ejecutan en paralelo para ahorrar tiempo. Por ejemplo, mientras se extraen datos puede que esté funcionando un proceso de transformación sobre los recibidos y de preparación para la carga, pero que empiece a funcionar un proceso de carga sobre los datos preparados, en lugar de tener que esperar a que termine todo el proceso de extracción (Microsoft, 2024).

Lecturas y material complementario

Video

Consulta el siguiente video para complementar esta introducción al diseño de procesos ETL:

- Análisis al Cuadrado
- ¿Qué son los procesos ETL? Curso ETL con Pentaho Gratis (Descripción) https://www.youtube.com/watch?v=jqZWYjubK3s

Tema 2. Integración de datos

Se trata de una estrategia clave que combina información proveniente de diversas fuentes para brindar a las organizaciones una perspectiva unificada. No es un proceso nuevo, pero

se ha vuelto más relevante con el creciente volumen de datos generado en la actualidad. Ha evolucionado de ser un proceso convencional a una actividad de vital importancia para las organizaciones que manejan grandes cantidades de datos. Sin embargo, a pesar de sonar simple, existen numerosos desafíos en el camino hacia una gestión y utilización exitosa de los datos.

Plataformas de redes sociales, dispositivos IoT, aplicaciones móviles y otros puntos de contacto digitales, contribuyen a la generación de datos, alcanzando millones de terabytes de datos cada día. Solo mediante la integración de datos se pueden combinar diferentes fragmentos de esta cantidad de datos, de manera que permita obtener ideas significativas y aprovechar todo su potencial.

No obstante, cada negocio es único y, por lo tanto, también lo son sus necesidades de integración de datos. Esto hace aún más importante comprender los diversos desafíos que presenta la integración de datos (Estuary, 2024).

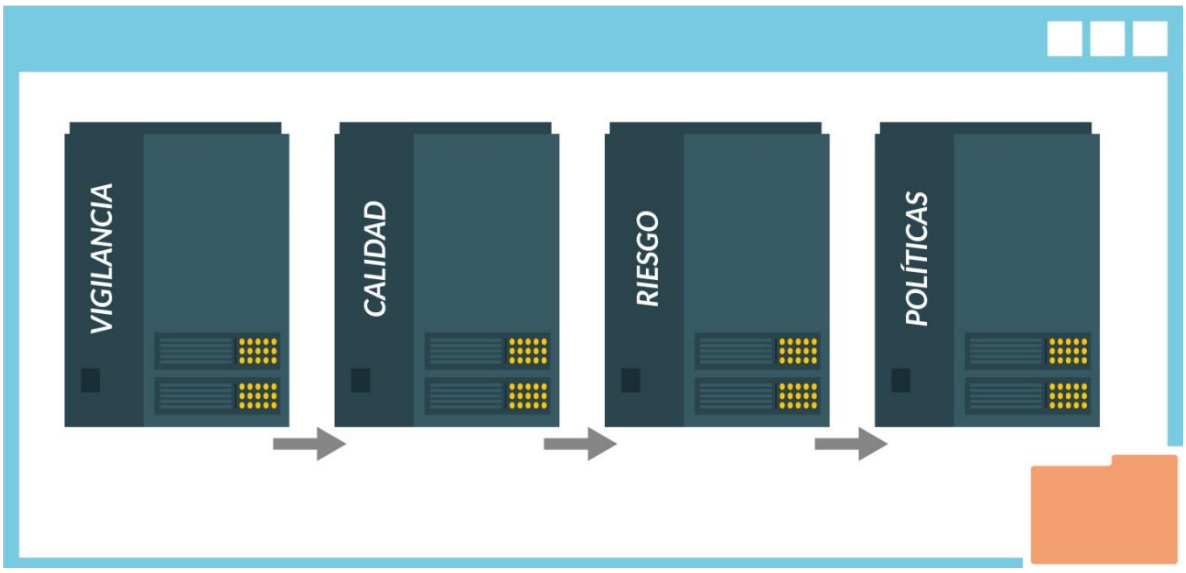
Desafíos de integración de datos:

1. **Establecimiento de un entendimiento común de los datos**

Aunque los datos son fundamentales en cualquier negocio, distintos equipos pueden interpretarlos y utilizarlos de manera diferente, lo que provoca inconsistencias. Para una integración de datos efectiva es necesario contar con un entendimiento común de los datos y un uso consistente en todos los equipos.

Crear un entendimiento compartido de los datos puede ser desafiante, requiere una comunicación clara, trabajo colaborativo y articulación entre los equipos, así como estándares de datos bien definidos y políticas de uso. Sin un entendimiento compartido de los datos la integración de datos se puede enfrentar a problemas, errores, ineficiencias y disputas entre equipos (Estuary, 2024).

Figura 2. Entendimiento común de datos



Fuente: adaptado de Estuary, (2024).

2. **Comprensión de los sistemas de origen y destino**

Comprender los sistemas de origen y destino es un desafío importante que enfrentan las organizaciones. El sistema de origen es la ubicación original donde se almacenan los datos, mientras que el sistema de destino es el destino al que deben transferirse los datos, como un almacén de datos.

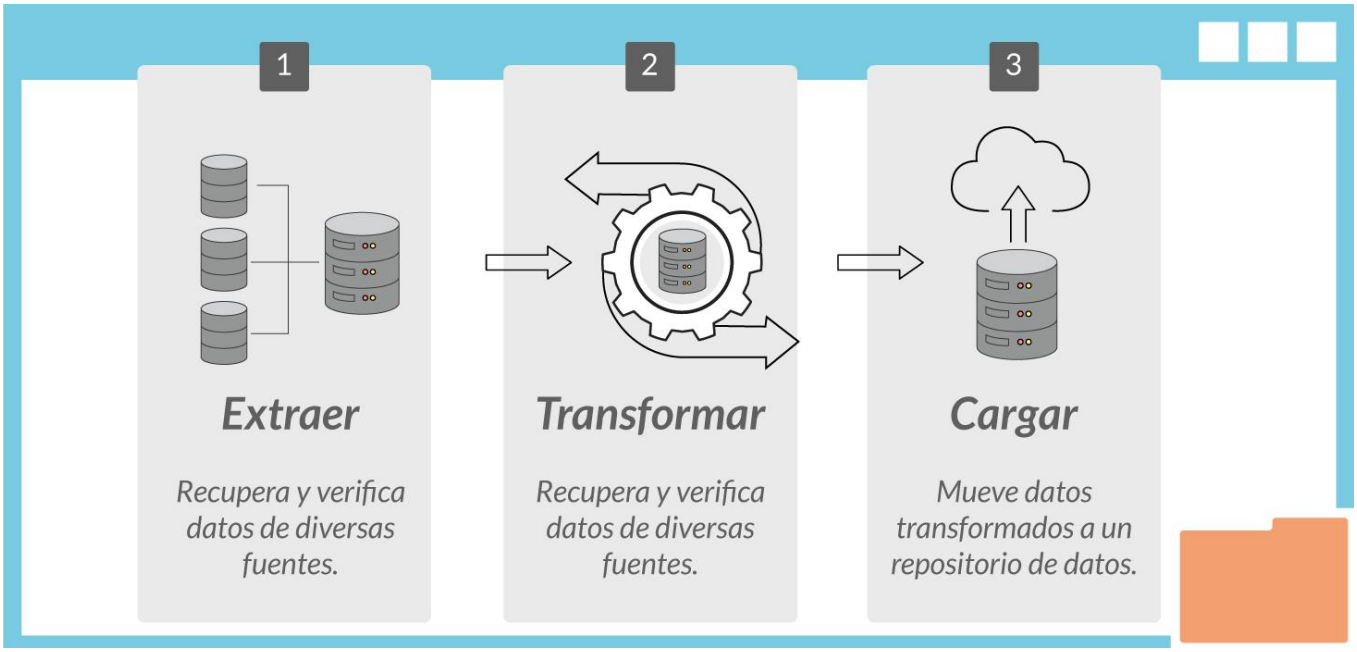
El desafío surge porque los datos pueden provenir de una amplia variedad de sistemas de almacenamiento de datos. Además, los datos en estos sistemas pueden cambiar a diferentes velocidades, lo que dificulta más el proceso de integración (Estuary, 2024).

3. **Mapeo de estructuras de datos heterogéneas**

Las estructuras de datos suelen diferir según los desarrollos individuales de los sistemas, cada uno con reglas únicas para almacenar y cambiar datos. Este desafío se vuelve aún más complejo cuando los sistemas utilizan diferentes formatos de datos, esquemas e idiomas. Convertir dichos datos en un formato uniforme para la integración puede consumir tiempo y provocar errores.

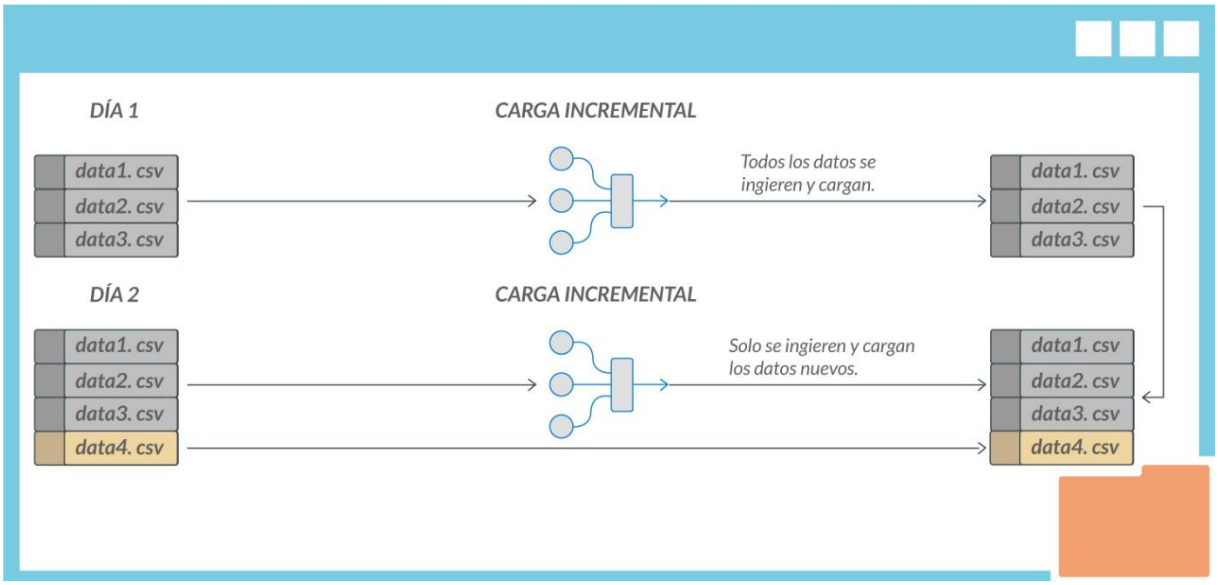
Sin un plan adecuado para manejar diversas estructuras de datos, existe un mayor riesgo de pérdida o corrupción de datos, lo que genera un análisis de datos inexacto y decisiones comerciales deficientes (Estuary, 2024).

Figura 3. Solución manejo de grandes volúmenes de datos



4. A medida que tu empresa continúa creando y recopilando más datos, el procesamiento e integración de esta información se vuelve compleja. Los grandes volúmenes de datos pueden abrumar los métodos tradicionales de integración de los mismos, provocando tiempos de procesamiento más largos y un mayor uso de recursos (Estuary, 2024).

Figura 4. Carga incremental



5. Gestión de infraestructura

Debes planificar de manera confiable la infraestructura de la integración de datos, con el fin de adelantarte a dificultades o enfrentar problemas como caídas del sistema, interrupciones de red o fallos de hardware que pueden interrumpir la integración y causar pérdida, retrasos o inexactitudes.

Además del hardware, la infraestructura de integración de datos también incluye herramientas, plataformas y sistemas de software para la extracción, transformación y carga de datos. Cambios o actualizaciones en estas herramientas, pueden afectar el proceso de integración y requerir optimización del flujo de trabajo (Estuary, 2024).

6. Gestión de costos inesperados

La complejidad de la integración de datos puede dar lugar a situaciones inesperadas y aumentar los costos. Estas situaciones pueden surgir por cambios en los datos o sistemas, después de establecer la integración. Por ejemplo, cambios en el formato o estructura de los datos requieren modificaciones en el flujo de trabajo, generando costos adicionales en tiempo y recursos.

Estos costos inesperados pueden sobrecargar tu presupuesto y recursos, además de afectar potencialmente a otros proyectos u operaciones. Si no se gestionan de manera efectiva, los costos restan valor a la integración de datos y la hacen menos rentable (Estuary, 2024).

7. **Accesibilidad de datos**

   Es la capacidad de acceder a los datos en la ubicación necesaria para la integración, independiente de dónde se almacenen o cómo se gestionan. Sin embargo, con la dispersión en diferentes sistemas, bases de datos y ubicaciones resulta difícil asegurar que todos los datos necesarios estén accesibles para la integración.
   La curación manual de datos depende en gran medida del esfuerzo humano. Esta dependencia provoca retrasos, inconsistencias y errores en el proceso de integración de datos. Asimismo, y a medida que aumenta el volumen y la complejidad de los datos, la tarea de curación de datos se vuelve más desafiante y consume más tiempo (Estuary, 2024).

8. **Gestión de calidad de datos**

   La alta calidad de los datos se caracteriza por ser precisa, consistente y confiable. No obstante, mantener una calidad de datos elevada se vuelve desafiante cuando estos provienen de distintas fuentes, cada una con sus propias reglas y formatos.
   La baja calidad de los datos puede originarse en errores de entrada, variaciones en los formatos o estructuras de datos, así como en datos desactualizados o faltantes. Dichos problemas conducen a la obtención de datos incorrectos, generando percepciones equivocadas y decisiones erróneas (Estuary, 2024)

Figura 5. Reglas para garantizar la calidad de datos

Fuente: adaptado de Estuary, (2024)

9. Seguridad y privacidad de los datos

Asegurarse de que los datos estén protegidos en todas las etapas del proceso de integración, es fundamental. Esto es especialmente importante al tratar con datos sensibles, como información personal, datos financieros o información comercial propietaria. Cualquier filtración o mal uso de datos puede dañar la reputación de la empresa o negocio, haciendo que los clientes y las partes interesadas pierdan confianza.

A medida que las amenazas cibernéticas continúan evolucionando y las regulaciones de protección de datos se vuelven más estrictas en muchas jurisdicciones, debemos asegurar que los procesos de integración de datos cumplan con las leyes y normas relevantes (Estuary, 2024).

10. Gestión de datos duplicados

Cuando los registros de datos aparecen más de una vez en el conjunto de datos se crean problemas durante la integración de estos. Los duplicados pueden originarse a partir de errores en la entrada de datos, errores del sistema, diferentes formatos o estructuras de datos.
Los datos duplicados hacen que los datos integrados sean inexactos y obstaculicen la toma de decisiones. Además, manejarlos consume mucho tiempo y recursos, lo que complica y encarece su integración (Estuary, 2024).

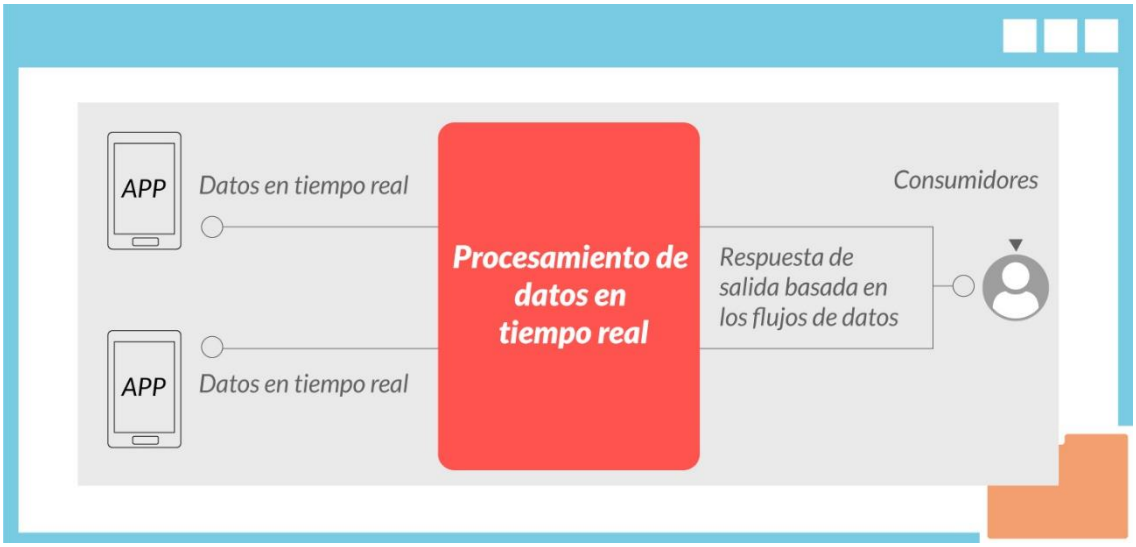
11. Optimización del rendimiento

Tu solución de integración de datos debe ser lo suficientemente capaz como para manejar grandes volúmenes de datos y tareas complejas sin afectar demasiado el rendimiento del sistema. A medida que aumenta el volumen y la complejidad de los datos, los procesos de integración de datos pueden volverse más intensivos en recursos y ralentizar los sistemas. Sin mejorar el rendimiento, la integración de datos puede ralentizar otros procesos, haciendo difícil que tu organización obtenga información oportuna y precisa (Estuary, 2024).

12. Integración de datos en tiempo real

Algunos procesos empresariales requieren la recolección de datos en tiempo real o casi en tiempo real, pero los retrasos pueden interrumpir dichos procesos. En ese sentido, integrar datos en tiempo real puede ser complejo, requiere la captura, procesamiento e integración de datos tan pronto como se generan, sin causar retrasos ni interrupciones.

Figura 6. Solución integración de datos en tiempo real



También es necesario gestionar el alto volumen y la velocidad de los datos, sin una estrategia efectiva de integración de datos para manejar datos en tiempo real será difícil obtener información oportuna, lo que afectará la toma de decisiones en la organización o negocio (Estuary, 2024).

Tema 3. Técnicas

Las **técnicas de elaboración de perfiles de datos** son métodos diversos que ayudan a analizar, evaluar y comprender los datos. Algunos de ellos son (Astera, 2024):

1. **Perfilado de columnas:**
   Analiza cada columna de una base de datos, observando el tipo de datos, la longitud y la presencia de valores vacíos. El análisis de frecuencia es crucial para identificar patrones y valores inusuales.

2. **Perfilado entre columnas:**
   Se centra en las relaciones entre diferentes columnas dentro de la misma tabla. Incluye análisis de claves y dependencias. El análisis de claves busca columnas con valores únicos, mientras que el análisis de dependencia examina cómo los valores de una columna dependen de los valores de otra. Esto ayuda a encontrar conexiones e inconsistencias.

3. **Perfilado entre tablas:**
   Analiza las relaciones entre diferentes tablas en una base de datos. Incluye el análisis de claves externas, el cual identifica columnas que coinciden con columnas de clave única en otra tabla. Esto revela cómo los datos de una tabla se relacionan con los de otra, proporcionando información sobre la estructura y precisión de la base de datos.

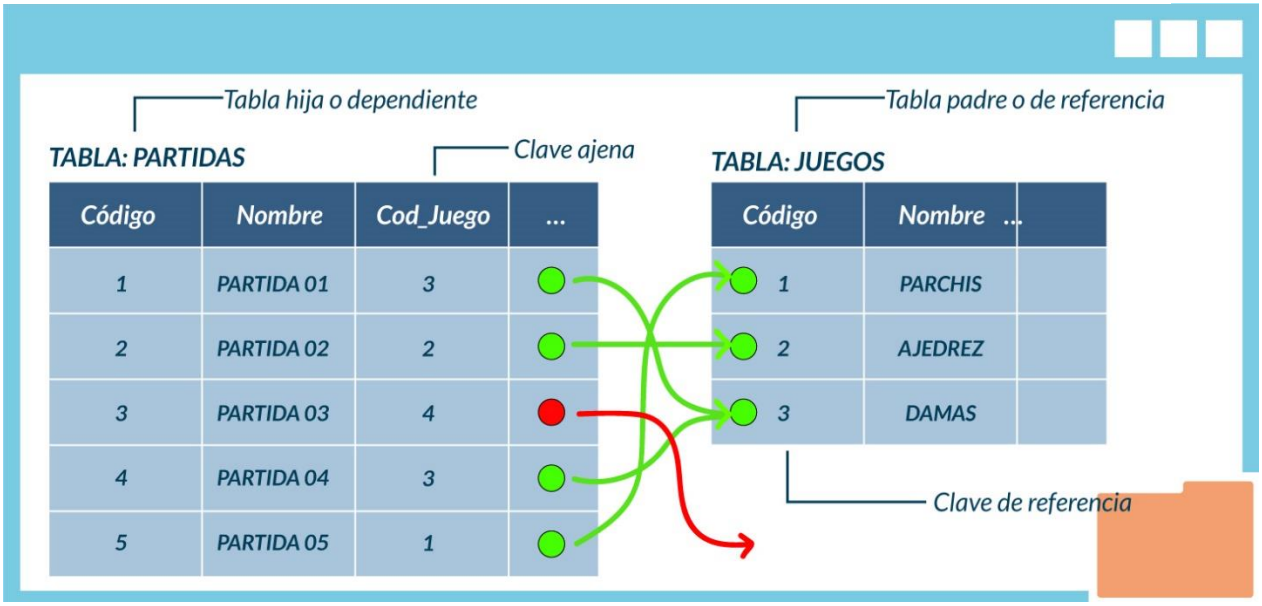
4. **Validación de datos:**
   Verifica la precisión y calidad de los datos según criterios o estándares específicos; incluye comprobaciones de formato, rango y coherencia para garantizar la limpieza, corrección y lógica de los datos.

Las reglas de validación en un proceso ETL (Extracción, Transformación y Carga)

Son criterios y condiciones que se aplican a los datos durante su movimiento desde la fuente hasta el destino. Estas reglas aseguran que los datos cumplan con los estándares y requisitos establecidos, antes de ser almacenados o utilizados para análisis. Aquí hay algunas reglas de validación comunes en un proceso ETL:

- **Integridad referencial:**
  Verifica que las relaciones entre las tablas se mantengan. Esto incluye la validación de claves primarias y foráneas, para garantizar así la coherencia en las relaciones entre las entidades.

Figura 7. Integridad referencial



- **Formato de datos:**

  Asegura que los datos cumplan con el formato especificado. Puede incluir la validación de fechas, números, códigos postales u otros formatos específicos según los requisitos del negocio.

- **Duplicados:**

  Identifica y maneja duplicados para evitar la redundancia de datos. Puede implicar la eliminación de duplicados o la consolidación de información.

- **Valores nulos:**

  Garantiza que los campos requeridos no estén vacíos y que se manejen adecuadamente los valores nulos según las políticas establecidas.

- **Reglas de negocio específicas:**

  Aplica reglas de negocio específicas de la organización, con el fin de garantizar que los datos cumplan con los requisitos y estándares del negocio.

- **Rango y límites:**

  Verifica que los valores caigan dentro de rangos predefinidos. Esto es crucial para garantizar la coherencia y la validez de los datos.

- **Consistencia temporal:**

  Para datos que cambian con el tiempo, se aplican reglas con el objetivo de garantizar tanto la consistencia temporal, como la vigencia de las relaciones o la versión correcta de los datos en un momento dado.

- **Conformidad con estándares:**

  Asegura que los datos cumplan con estándares específicos de la industria o regulaciones gubernamentales como GDPR (Reglamento General de Protección de Datos) u otros requisitos de cumplimiento.

- **Validación de reglas de transformación:**

  Si hay reglas de transformación durante el proceso ETL, se deben validar para tener certeza de que los datos transformados sean coherentes y precisos.

Monitoreo de cargas:

- Implementa mecanismos para monitorear y registrar estadísticas sobre el proceso ETL, identificando posibles problemas o desviaciones en el flujo de datos.

Estas reglas de validación son esenciales para mantener la calidad y la integridad de los datos a lo largo del proceso ETL, garantizando que los datos resultantes sean confiables y útiles en la toma de decisiones.
¡Has finalizado la tercera unidad del curso! En esta última parte nos enfocamos en el diseño de procesos ETL. Esperamos que identifiques la necesidad de habilidades para organizar la transformación de datos y el impacto que estos procesos traen a la toma de

decisiones. Aprendiste sobre integración de datos y técnicas para la gestión de datos, siempre buscando la coherencia de la información.

Esperamos que, con los temas estudiados, los ejemplos y recursos recomendados, te acerques cada vez más a una eficiente y correcta manera de administrar, gestionar y manipular información en bases de datos.

Referencias bibliográficas

- Análisis al Cuadrado, (6 de julio de 2019). ¿Qué son los procesos ETL? Curso ETL con Pentaho Gratis (Descripción). [Video]. YouTube. https://www.youtube.com/watch?v=jqZWYjubK3s

- Astera. (2024). Elaboración de perfiles de datos: tipos, técnicas y mejores prácticas | Astera. https://www.astera.com/es/type/blog/data-profiling/

- Datalized, (29 de diciembre de 2021). Paso 1: Integración de datos. [Video]. YouTube. https://www.youtube.com/watch?v=db6XO6XSld0

- Estuary. (2024). 11+ Most Common Data Integration Challenges & Solutions | Estuary. https://estuary.dev/data-integration-challenges/

- Johnny Ahumada Pereira, (11 dic 2020). Técnicas de integración y Visualización de Datos Master Class #1. [Video]. YouTube. https://www.youtube.com/watch?v=Si37GvsvWCE

- Microsoft. (2024). Extracción, transformación y carga de datos (ETL) - Azure Architecture Center | Microsoft Learn. https://learn.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl

IU Digital de Antioquia

INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA

www.**iudigital**.edu.co