|  |  |
| --- | --- |
| PROGRAMA DE FORMACIÓN | Gestión analítica de datos |

|  |  |  |  |
| --- | --- | --- | --- |
| COMPETENCIA | 220501113 - Administrar la base de datos de acuerdo con los estándares y requisitos técnicos. | RESULTADOS DE APRENDIZAJE | 220501113-03 -Controlar la gestión de datos de acuerdo con especificaciones técnicas. |

|  |  |
| --- | --- |
| NÚMERO DEL COMPONENTE FORMATIVO | CF15 |
| NOMBRE DEL COMPONENTE FORMATIVO | Procesos ETL (extracción, transformación y carga) |
| BREVE DESCRIPCIÓN | Los procesos de extracción, transformación y carga se utilizan en los proyectos de ciencias de datos para extraer información de diversas fuentes, transformarla para adaptarla a los requerimientos del proyecto en desarrollo y cargarla en un espacio compartido, donde los *stakeholders* de la organización puedan emplearla útilmente. Esto posibilita un mejor aprovechamiento de los datos disponibles. |
| PALABRAS CLAVE | Extracción, transformación de datos, *data warehouse*, minería de datos, *web scraping* |

|  |  |
| --- | --- |
| ÁREA OCUPACIONAL | Ciencias naturales, aplicadas y relacionadas |
| IDIOMA | Español |

# Tabla de contenidos

**Introducción**

**1. ¿Qué es ETL?**

**2. Extracción de datos**

2.1. Análisis de necesidades de información

2.2. Técnicas de extracción de datos

2.3. Minería de datos: conceptos y herramientas

**3. Transformación de datos**

3.1. Técnicas de limpieza y transformación de datos

3.2. Herramientas para transformación y limpieza de datos

**4. Carga de datos**

4.1. *Data warehouse*: conceptos y herramientas

4.2. Tecnologías de almacenamiento (OLAP, ROLAP, MOLAP, HOLAP)

4.3. Mantenimiento de *data warehouse*

# Introducción

|  |
| --- |
| Cuadro de texto |
| Estimado aprendiz, reciba un cordial saludo y sea bienvenido a este componente formativo en el que conocerá las actividades realizadas en los procesos de extracción, transformación y carga de datos en un proyecto de ciencia de datos. En este componente aprenderá técnicas para la extracción de información de diversas fuentes, así como recomendaciones para realizar la limpieza e imputación de datos, todo esto con el objetivo de preparar la información disponible y aprovecharla al máximo para el beneficio de la organización.  El enlace que encontrará a continuación lo dirigirá al video introductorio, con el cual podrá conocer sobre lo que trabajará durante esta etapa formativa. Deseamos que pueda culminar satisfactoriamente este proceso. ¡Muchos éxitos! |

**Guion de video introductorio**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tipo de recurso** | Video spot animado | | | |
| **NOTA** | **La totalidad del texto locutado para el video no debe superar las 390 palabras aproximadamente** | | | |
| **Título** | Introducción a los procesos de extracción, transformación y carga | | | |
| **Escena** | **Imagen** | **Sonido** | **Narración (voz en off)** | **Texto** |
| **1** | Imagen que muestre a una persona en un PC haciendo recolección de datos.  Hand using laptop computer with virtual screen and document for online approve paperless quality assurance and erp management concept  Fuente: <https://www.freepik.com/free-photo/hand-using-laptop-computer-with-virtual-screen-document-online-approve-paperless-quality-assurance-erp-management-concept_24755711.htm> | Música de fondo | Uno de los principales retos que enfrenta una organización es cómo debe realizar los procesos de recolección de datos de las diversas fuentes que alimentan el sistema de información, saber dónde estará ubicada la misma para facilitar su acceso y uso y conocer si los datos recolectados se encuentran en un formato útil y claro para su aprovechamiento. | La extracción de datos, su transformación y carga representan un reto para la organización. |
| **2** | Imagen que muestre a un analista de datos en un PC.  Cloud computing data management concept  Fuente: <https://img.freepik.com/free-photo/cloud-computing-data-management-concept_53876-127747.jpg?w=740&t=st=1666582477~exp=1666583077~hmac=4c501dba23213d30c213fec594324809b760c1d2eac135386c5229fe047830d3> | Música de fondo | Para afrontar estos retos, se definieron procesos de extracción, transformación y carga (ETL, por sus siglas en inglés), que orientan a los científicos de datos, analistas e investigadores en cómo canalizar la información desde su recopilación de diversos orígenes, pasando por su transformación, según las necesidades y hasta su posterior almacenamiento, con el fin de que la información pueda ser explotada por la organización. | Extracción  Transformación  Carga |
| **3** | Imagen que muestre extracción de datos    Fuente:<https://learn.microsoft.com/es-es/azure/architecture/data-guide/images/etl.png> | Música de fondo | En los procesos de extracción se obtienen los datos de las fuentes internas y externas que están disponibles. Estos datos pueden tener diferentes orígenes y formatos, como hojas de cálculo, bases de datos y archivos planos.  Por su parte, en el proceso de transformación se convierten los datos desde su formato inicial al requerido para su utilización. Esta fase incluye actividades de transformación y de limpieza y su objetivo es mejorar la calidad de los datos extraídos de diversas fuentes, mediante la corrección de inconsistencias, imprecisiones y pérdida de valores.  Una vez extraídos y transformados, los datos deben cargarse en los contenedores o *data* *warehouse* para que estén disponibles y que el equipo de trabajo pueda utilizarlos en las aplicaciones requeridas o en los proyectos de ciencias de datos de la organización. | Extracción: se toman datos de diversas fuentes de información y formatos.  Transformación: se limpia y depura la información extraída.  Carga: hacer que la información transformada esté disponible para su uso y aprovechamiento. |
| **4** | Imagen que muestre proceso de extracción, transformación o carga de datos.  Document management system dms .businessman hold folder and document icon.software for archiving, searching and managing corporate files and information.internet technology concept.digital security  Fuente: <https://img.freepik.com/premium-photo/document-management-system-dms-businessman-hold-folder-document-icon-software-archiving-searching-managing-corporate-files-information-internet-technology-concept-digital-security_150455-9195.jpg?w=740> | Música de fondo | De la correcta realización de las actividades de extracción, transformación y carga dependerá el éxito o fracaso del proceso de reporte, análisis, *machine learning* e inteligencia artificial que requiera implementar la organización. | Importante para procesos como   * Reporte * Análisis * *Machine learning* * Inteligencia artificial |
| **5** | Imagen que muestre a una persona en un PC trabajando con datos.  Document management system dms online documentation database and process automation to efficiently manage files knowledge and documentation in enterprise with erp corporate business technology  Fuente: <https://img.freepik.com/premium-photo/document-management-system-dms-online-documentation-database-process-automation-efficiently-manage-files-knowledge-documentation-enterprise-with-erp-corporate-business-technology_265022-7253.jpg?w=740> | Música de fondo | Algunas de las actividades requeridas en la transformación de datos implican realizar operaciones de filtrado, agregación, combinación (generación de nuevas variables), limpieza y validación de datos (valores y formatos), eliminación de duplicado y manejo de valores nulos, que puedan afectar el uso de la información. | Ejemplos de actividades de transformación:   * Agregación * Combinación * Limpieza * Validación * Imputación |
| **6** | Imagen que muestre datos en la nube.  Cloud storage banner background  Fuente: <https://img.freepik.com/free-photo/cloud-storage-banner-background_53876-108506.jpg?w=740&t=st=1666582825~exp=1666583425~hmac=df74f59d838620ffbddc36ad59961e22ba5c3801017836e37be9b9b0e60bb77f> | Música de fondo | En muchas organizaciones, la información y los conjuntos de datos a analizar se encuentran almacenados en la nube y, al igual que con los repositorios de información tradicionales, es necesario centralizar los datos requeridos y ajustar sus formatos para que sean más útiles y fáciles de manejar. La computación en la nube y las herramientas que se están desarrollando buscan hacer estos procesos cada vez más eficientes y lograr con ello, que el principal foco de los científicos de datos sea el análisis de la información y no su preparación. | La computación en la nube y las herramientas para procesos de ETL facilitan el trabajo de los científicos y los analistas de datos. |
| **7** | Imagen que muestre el proceso de extracción, transformación y carga de datos.  Digital increasing bar graph with businessman hand overlay  Fuente: <https://img.freepik.com/free-photo/digital-increasing-bar-graph-with-businessman-hand-overlay_53876-97640.jpg?w=900&t=st=1666582878~exp=1666583478~hmac=2aafc067a0556a6bbcd215babd2370ab8186bda6625e73cafe85cda806061e16> | Música de fondo | En el transcurso del contenido, el aprendiz conocerá las diversas técnicas y herramientas disponibles para la realización de los procesos de ETL, como también recomendaciones y buenas prácticas para optimizar la ejecución de estas actividades. Los videotutoriales de herramientas y las actividades ayudarán a validar que el aprendiz adquiera el conocimiento transmitido y pueda aplicarlo en un ejercicio propio de ciencia de datos para su organización, con los mejores resultados. | Técnicas  Herramientas  Buenas prácticas |
| **Nombre del archivo** | 228131\_V\_01 | | | |

**Desarrollo de contenidos**

# 1. ¿Qué es ETL?

|  |  |
| --- | --- |
| Tipo de recurso | Cajón de texto de color |
| Los procesos *extract*, *transform* *and* *load* (ETL), en español, extracción, transformación y carga, son parte esencial de los procesos de ciencia de datos, en los que se aplican diferentes herramientas, técnicas y actividades que permiten, como su nombre lo indica, *extraer* la información de las diversas fuentes con que cuenta la organización (estas pueden estar soportadas en sistemas digitales o físicos). Luego, está *transformar* esa información mediante procesos de limpieza, validación e imputación, que permitan tener un conjunto de datos de calidad que sea íntegro, seguro y utilizable. Por último, se debe *cargar* la información en un repositorio central, donde la organización pueda aprovecharla. La secuencia de aplicación de estas fases se observa con mayor claridad a continuación:    Fuente: <https://learn.microsoft.com/es-es/azure/architecture/data-guide/images/etl.png>  **Imagen:** 228131\_i\_01  De la calidad en la realización de estas tareas dependerá en gran medida el éxito o fracaso de los ejercicios de ciencias de datos. En este ámbito de uso y aprovechamiento de la información, se dice de manera constante que, “si entra basura, sale basura”. Si los datos y la información recolectada son incompletos, incorrectos y no son claros, los resultados y los análisis realizados con dicha información serán de poca o nula utilidad para la organización. Es por esto que en los últimos años se ha trabajado arduamente en el desarrollo de mejores herramientas para apoyar los procesos de ETL, así como procesos más claros y metodologías y marcos de trabajo que permitan estandarizar y hacer más eficiente esta labor. | |

# 2. Extracción de datos

|  |  |  |
| --- | --- | --- |
| **Tipo de recurso** | Slider pasos | |
| **Introducción** | El proceso de extracción de datos consiste en recopilar desde las diversas fuentes la información requerida para la realización del proyecto de ciencias de datos. Estas fuentes son, en la mayoría de casos, muy diversas, dado que incluyen archivos físicos, como facturas, registros en papel, libros contables y respaldo, o también digitales, como hojas de cálculo, documentos de texto, bases de datos, entre otros. Los formatos y la compatibilidad entre estas fuentes de información hacen que la extracción se convierta en un reto importante para la organización.  En algunos casos, al revisar los datos requeridos para el proyecto de ciencias de datos, se puede encontrar que la organización no cuenta con la información requerida para el proceso, ante lo cual es necesario definir los elementos faltantes y su posible fuente y diseñar instrumentos para su levantamiento.  A continuación, se presentan los pasos a realizar para un correcto proceso de extracción de datos: | |
| **Slide 1** | Definir o identificar los datos requeridos para el proyecto de ciencia de datos y su ubicación origen. | Ciencias informáticas forenses. análisis de evidencia digital, investigación de delitos informáticos, recuperación de datos. experto en ciberseguridad identificando actividad fraudulenta. ilustración de metáfora de concepto aislado de vector  Fuente: <https://www.freepik.es/vector-gratis/ciencias-informaticas-forenses-analisis-evidencia-digital-investigacion-delitos-informaticos-recuperacion-datos-experto-ciberseguridad-identificando-actividad-fraudulenta-ilustracion-metafora-concepto-aislado-vector_12083574.htm#query=identificar%20datos&position=9&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_02 |
| **Slide 2** | Extraer los datos desde su fuente de origen. | El científico de datos y el especialista extraen conocimientos y perspectivas de los datos. análisis de ciencia de datos, control de aprendizaje automático, concepto de análisis de big data. ilustración aislada de bluevector coral rosado  Fuente: <https://www.freepik.es/vector-gratis/cientifico-datos-especialista-extraen-conocimientos-perspectivas-datos-analisis-ciencia-datos-control-aprendizaje-automatico-concepto-analisis-big-data-ilustracion-aislada-bluevector-coral-rosado_11668808.htm#query=extraer%20datos&position=0&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_03 |
| **Slide 3** | Analizar la calidad y el estado de los datos extraídos. | Analizar la ilustración del concepto  Fuente: <https://www.freepik.es/vector-gratis/analizar-ilustracion-concepto_15110186.htm#query=analizar%20datos&position=2&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_04 |
| **Slide 4** | Definir a partir de análisis de la calidad realizado si los datos extraídos cumplen con los requerimientos esperados. Si estos no cumplen, deberán ser rechazados. | Hombre revisando una lista de tareas gigante  Fuente: <https://www.freepik.es/vector-gratis/hombre-revisando-lista-tareas-gigante_4105833.htm#query=verificar&position=4&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_05 |

|  |
| --- |
| Cuadro de texto |
| Una acción para considerar en los procesos de extracción de datos es generar el menor impacto en el sistema de origen, dado que si se afecta el sistema original, esto mismo podría pasar con la actividad cotidiana de la organización. Para evitar este tipo de inconvenientes, se recomienda seleccionar los horarios para hacer el proceso de extracción o realizarlo en bloques o lotes, y así minimizar el impacto. |

## 2.1. Análisis de necesidades de información

|  |  |  |
| --- | --- | --- |
| **Tipo de recurso** | Tarjetas Avatar | |
| **Introducción** | En esta fase de extracción es muy importante tener claros los requisitos y expectativas de la organización, así como los objetivos del proceso de ciencia de datos que se realizarán, con el fin de dar respuesta y definir las necesidades de información. Estas necesidades son aquellos datos que el equipo debe tener para hacer a cabalidad su trabajo, investigación, estudio o predicción. Se puede dar el caso de que la organización necesite información que sabe o no que existe o que no sabe que necesita.  Un aspecto útil para identificar las necesidades y posibles fuentes de información es conocer cómo esta fluye al interior de la organización. En este sentido, la información fluye de diferentes maneras: | |
| De forma vertical ascendente. | |  |
| De subordinado a superior o vertical descendente. | |  |
| De superior a subordinado. | |  |
| De forma horizontal entre personas de un mismo nivel jerárquico, que tienen responsabilidades compartidas. | |  |

|  |  |
| --- | --- |
| Tipo de recurso | Cajón de texto de color |
| En esta fase también existen flujos externos de información con actores del entorno de la organización que deben ser identificados, si son necesarios para el proyecto. Todos estos flujos tienen que analizarse para identificar en ellos la información requerida para el proceso de ciencia de datos que se ejecutará y, por consiguiente, definirán cuál es la fuente de datos más pertinente y adecuada. En esta etapa también se analiza el tipo de información que se está generando, quién la usa, cómo y para qué y quiénes no deben tener acceso a esa información y por qué, con el fin de evitar compartir información delicada o confidencial con actores que no están autorizados.  Para modelar el flujo de la información a través de la organización, se utilizan los diagramas de flujo, los cuales permiten representar el proceso y el orden en que se ejecutan las tareas. El organigrama, los manuales de procedimiento y las entrevistas a actores claves serán un insumo fundamental para la elaboración de estos modelos.  ejemplo de diagrama de actividades  Fuente: <https://d2slcw3kip6qmk.cloudfront.net/marketing/pages/chart/activity-diagram-for-login-UML/activity-diagram-for-login-UML-650x797.png>  **Imagen:** 228131\_i\_06  Para hacer el análisis de necesidades de información, se usan dos enfoques, uno directo y otro indirecto. En el directo se utilizan técnicas para obtener de primera mano información del usuario. Algunas de estas técnicas son: entrevistas, cuestionarios, observación de hábitos de búsqueda y uso de información, técnica de incidente crítico y métodos de consenso (algunos de estos métodos utilizados son la técnica Delphi y los *focus groups* o grupos orientados). Por su parte, el enfoque indirecto consiste en el análisis de documentos, referencias y solicitudes de documentos, además de la observación y del registro de hechos. | |

|  |  |
| --- | --- |
| **Tipo de recurso** | Infografía estática |
| **Texto introductorio** | La siguiente infografía presenta algunas de las técnicas utilizadas en el proceso de análisis de necesidades de información: |
| **Figura 1**  *Técnicas para el análisis de necesidades de información* | |
| **Código de la imagen** | 228131\_i\_07 |

## 2.2. Técnicas de extracción de datos

|  |
| --- |
| Cuadro de texto |
| El proceso de extracción de datos se basa en acceder a fuentes de distintos orígenes para luego normalizar los datos adquiridos y almacenarlos para su uso por otras herramientas. Dependiendo de los formatos y de las diferentes fuentes, se pueden usar herramientas que faciliten y automaticen la consolidación de la información. Así que es importante tener en cuenta que dichas fuentes suelen estar mal organizadas y, en muchos casos, desestructuradas.  La extracción de datos se puede realizar de varias formas: manual, en la cual se requiere todo el tiempo de la intervención humana, automatizada y, por último, *human in the loop*, que es una combinación de las dos primeras. En la actualidad, y aprovechando los procesos de inteligencia artificial y las herramientas que se están diseñando, se busca que la extracción de datos sea automatizada, debido a algunas ventajas, como mejora de la precisión, reducción de costos, ahorro de tiempo y mejores tiempos de respuesta. |

|  |  |  |
| --- | --- | --- |
| **Tipo de recurso** | Slider Hitos/ Línea de tiempo horizontal | |
| **Introducción** | Uno de los procesos de extracción de datos es el que se hace de documentos físicos. Este se puede realizar siguiendo los pasos descritos a continuación: | |
| **Cargar el documento:** | El documento físico (en papel) debe ser transformado en un documento digital mediante herramientas como escáner o teléfono inteligente. Este documento digital, por lo general, se almacena en formatos como JPG, PDF o PNG, y se recomienda que su calidad sea óptima para facilitar los procesos de reconocimiento y análisis posteriores (imágenes de baja calidad pueden dificultar el proceso de reconocimiento y generar resultados erróneos). | Ilustración de redes sociales  Fuente: <https://www.freepik.es/vector-gratis/ilustracion-redes-sociales_5275573.htm#query=documento&position=13&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_08 |
| **Convertir documento a texto:** | Las tecnologías de OCR (reconocimiento óptico de caracteres u *optical character recognition*, en inglés) realizan la conversión electrónica o mecánica de imágenes de texto mecanografiado, manuscrito o impreso en texto codificado por la máquina, ya sea a partir de un documento escaneado, de una foto de un documento, de una foto de escena o de un texto de subtítulos superpuesto a una imagen. De esta forma, el texto o los datos incluidos en el documento físico podrán ser usados por aplicaciones para su gestión y almacenamiento. | Tome fotografías de recibos y gastos con reconocimiento óptico de caracteres (ocr) para reembolso  Fuente: <https://www.freepik.es/vector-premium/tome-fotografias-recibos-gastos-reconocimiento-optico-caracteres-ocr-reembolso_32682215.htm#query=tecnologia%20ocr&position=4&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_09 |
| **Análisis de sintaxis:** | Las herramientas de análisis de sintaxis permiten convertir el documento digital (ahora en texto) identificando las secciones que lo componen y pasándolo a un formato estructurado (JSON, XML, CSV). | Source code programming on screen of computer  Fuente: <https://www.shutterstock.com/es/image-photo/source-code-programming-on-screen-computer-1284744910>  Nota:  **Imagen:** 228131\_i\_10 |
| **Verificación de datos:** | Se recomienda realizar una verificación de los datos extraídos para validar su calidad. Según la cantidad de información, este proceso se puede hacer a todos los documentos convertidos o por medio de una muestra significativa. | Ilustración del concepto de codificación manual  Fuente: <https://www.freepik.es/vector-gratis/ilustracion-concepto-codificacion-manual_21532468.htm#query=verificaci%C3%B3n%20de%20datos&position=3&from_view=search&track=sph>    Nota:  **Imagen:** 228131\_i\_11 |
| ***scraping* o *web scraping*:** | También llamada *extracción de datos de la web*, consiste en desarrollar algoritmos o usar herramientas para recuperar documentos e información de una página web. El *scraping* hace uso del lenguaje de marcado (etiquetas HTML), con el cual están construidos los sitios web y expresiones regulares, para extraer de estos sitios solo la información requerida. Algunas de estas herramientas son: Octoparse, Cyotek WebCopy y Scraper (esta última funciona como una extensión del navegador Google Chrome). | Raspado web o recolección de información de un diagrama de esquema de datos de sitios web. Extracción de información digital educativa etiquetada a partir de la ilustración vectorial de sitios HTML. Proceso automático de recopilación de contenido de red  Fuente: <https://www.shutterstock.com/es/image-vector/web-scraping-information-harvesting-websites-data-2088809149>    Nota:  **Imagen:** 228131\_i\_12 |

|  |  |
| --- | --- |
| Tipo de recurso | Cajón de texto de color |
| Llamado a la acción  En el sitio web <https://tools.digitalmethods.net/> el aprendiz encontrará varias herramientas de *scraping*, análisis y recolección de datos, que podrá usar para recopilar información de diferentes fuentes digitales en Internet, como buscadores, redes sociales y repositorios. Por ejemplo, la herramienta YouTube Data Tools sirve para extraer información desde YouTube, usando la API de la plataforma. Entre los datos que se pueden tomar, se encuentra la información de un canal de videos específico o los comentarios de los videos de este canal. Este tipo de información puede ser muy útil al momento de realizar un estudio del grado de aceptación de una marca o empresa.  La consulta de ejemplo que se presenta en la siguiente imagen arroja como resultados archivos CSV con la información básica del video consultado, los comentarios, los autores y la red de comentarios.  **Figura 2**  *Interfaz herramienta YouTube Data Tools*      **Imagen:** 228131\_i\_13 | |

## 2.3. Minería de datos: conceptos y herramientas

|  |  |
| --- | --- |
| Tipo de recurso | Cajón de texto de color |
| La minería de datos es el proceso de búsqueda y análisis de información valiosa, que se utiliza en la toma de decisiones. Esta forma de obtener información se realiza en enormes cantidades de fuentes que manejan las organizaciones. La analogía con el concepto de minería se emplea por su similitud en la búsqueda de riquezas y materiales valiosos.  Sumado a lo anterior, la minería de datos es un proceso que utiliza técnicas matemáticas, estadísticas y de aprendizaje de máquina para identificar y extraer información útil de enormes cantidades de datos que existen, para luego convertirla en conocimiento en forma de patrones, (Rebollo, 2017). Por ejemplo, se puede utilizar para procesar grandes cantidades de información de clientes e identificar sus hábitos de compra, y así poder definir nuevos productos y características o hacer mejoras en los productos que ya están en circulación.  Este proceso se enfoca en realizar dos operaciones:   * Predecir tendencias y comportamientos * Identificar patrones previamente desconocidos   Un ejemplo de la aplicación de la minería de datos para la identificación de patrones es la detección de transacciones fraudulentas con tarjetas de crédito. Conforme se utiliza una tarjeta de crédito de manera normal y periódica, se identifican en estos usos patrones de hábitos de compra, lugares de uso y cantidades de dinero retiradas o gastadas en compras. Cuando esta tarjeta se roba o se pierde y se emplea de forma fraudulenta, su nuevo uso tendrá un patrón diferente al acostumbrado, lo cual generará alertas al cliente o bloqueos por seguridad.  **Herramientas para la minería de datos:** existen diversas metodologías y marcos de trabajo para los proyectos de minería de datos. Uno de los modelos más utilizados y aceptados por investigadores, científicos de datos, desarrolladores y usuarios es CRISP-DM (*Cross Industry Standard Process for Data Mining*), que fue propuesto en los años noventa por un consorcio europeo de empresas. Esta metodología plantea un ciclo de vida compuesto por seis etapas secuenciales (se ejecuta una después de la otra), donde es común la retroalimentación y el retorno a fases previas según se avanza en el proceso. | |

|  |  |
| --- | --- |
| **Tipo de recurso** | Infografía estática |
| **Texto introductorio** | Como ya se mencionó, existen seis etapas secuenciales en el ciclo de vida del modelo CRISP-DM. Estas se ilustran a continuación, de la siguiente manera: |
| Ciclo de vida de minería de datos  Fuente: <https://www.ibm.com/docs/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/images/crisp_process.jpg>  Nota: | |
| **Código de la imagen** | 228131\_i\_14 |

|  |  |
| --- | --- |
| Tipo de recurso | Cajón de texto de color |
| **Llamado a la acción**  Si desea conocer más herramientas tanto de *software* propietario (pagas) como de código abierto, que se pueden emplear para los procesos de minería de datos, visite los siguientes enlaces:  SPSS: <https://www.ibm.com/products/spss-statistics>  SAS: <https://www.sas.com/en_us/home.html>  Microsoft SQL Data Mining: <https://www.microsoft.com/es-mx/evalcenter/evaluate-sql-server-2019>  R: <https://www.r-project.org/>  KNIME: <https://www.knime.com/> | |

# 3. Transformación de datos

|  |
| --- |
| Cuadro de texto |
| En esta fase se realiza la transformación de los datos corrigiendo y resolviendo posibles inconvenientes que pueda presentar la información para que la organización la aproveche. Esto se hace a través de un conjunto de reglas que buscan proporcionar orden y claridad a la información recopilada. Las reglas y los parámetros que debe cumplir la información dependerán de los criterios y necesidades de cada organización.  En esta etapa se realizan actividades de validación, eliminación de duplicados, codificación, ajustes de formato y otras requeridas, que ayudan a identificar qué datos son útiles y cuáles son deficientes para iniciar el proceso. En el caso de los datos deficientes, es necesario decidir qué se hará con ellos, si se omiten o se ajustan de alguna manera para realizar el análisis de la información. |

## 3.1. Técnicas de limpieza y transformación de datos

|  |  |
| --- | --- |
| Tipo de recurso | Cajón de texto de color |
| En la actualidad, los datos son el principal activo de las organizaciones y el insumo principal en los procesos de ciencias de datos e inteligencia de negocios, así que garantizar su calidad es una de las actividades más importantes de estos procesos. La limpieza de estos datos, también llamada *data cleansing* o *data scrubbing*, consiste en resolver inconsistencias en los conjuntos de datos. Esto incluye corregir o eliminar registros inexactos o incompletos, eliminar información irrelevante, ajustar los formatos de los datos, entre otros. Cuando la información proviene de diferentes fuentes, aumenta la necesidad de realizar un adecuado proceso de limpieza de datos, debido a que diversas fuentes pueden contener formatos de datos redundantes, incompatibles o sucios.  Las siguientes son algunas de las ventajas de la limpieza de datos:   * Mejor calidad en los datos, que repercute de forma positiva en todos los procesos y actividades realizados con la información. * Permite cumplir con normatividades vigentes, relacionadas con la privacidad y uso de los datos. * Uso de datos de mejor calidad, que puede ayudar a mejorar los procesos internos de la organización. | |

|  |  |
| --- | --- |
| **Tipo de recurso** | Acordeón tipo 1 |
| **Introducción** | Las técnicas de limpieza de datos pueden variar según el tipo de información y de organización. Para realizar de este proceso, se describen algunos pasos básicos a continuación: |
| Borrar concepto. Pequeñas personas borrando datos en el portátil y trasladando archivos innecesarios al basurero. Limpiando la memoria digital. Estilo moderno de dibujos animados planos. Ilustración del vector sobre fondo blanco  Fuente: <https://www.shutterstock.com/es/image-vector/delete-concept-tiny-people-deleting-data-1901156443>  Nota:  **Imagen:** 228131\_i\_15 | |
| **Eliminación de datos duplicados o irrelevantes**: cuando se recopila información desde diferentes fuentes, existe una alta probabilidad de obtener datos duplicados. En este caso, se deben eliminar todos los datos que se repiten o combinar estos registros para no perder información (en algunos casos, los registros duplicados poseen cada uno información que el otro no). También se deberá eliminar cualquier otro tipo de dato que no sea relevante para el objetivo del proyecto de ciencias de datos. | |
| **Corrección de errores estructurales**: esto ocurre cuando en los conjuntos de datos se encuentran errores tipográficos o nomenclaturas extrañas, lo que puede ocasionar categorías mal etiquetadas. Para esto, se pueden usar filtros o funciones (según las herramientas que se estén utilizando) para eliminar o corregir los registros con estas características. | |
| **Tratamiento de valores atípicos *(outliers)*:** este tipo de datos se debe administrar con mucha atención, porque los valores atípicos no necesariamente son valores incorrectos. En caso de ser incorrectos, se pueden reemplazar por valores como el promedio. En otros casos, estas observaciones únicas se eliminan para que no alteren los cálculos realizados con los valores del atributo. | |
| **Manejo de datos faltantes (imputación):** un factor a considerar es no ignorar los datos faltantes en los conjuntos de datos, dado que muchos procedimientos, algoritmos y funciones no aceptan este tipo de valores y pueden arrojar errores y resultados incorrectos. Existen varias opciones para tratar este tipo de problemas. La primera es eliminar los registros con valores perdidos, y solo se recomienda cuando son pocos los que tienen estos problemas. La segunda consiste en ingresar valores perdidos con base en otros registros. Esta opción implica analizar los registros existentes, similares a los que tienen problemas, con el fin de identificar el valor posible que puede asignarse al dato faltante. La tercera opción es modificar la forma en que se usan los datos, para usar los demás de manera efectiva. | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tipo de recurso** | Video spot animado | | | |
| **NOTA** | **La totalidad del texto locutado para el video no debe superar las 500 palabras aproximadamente** | | | |
| **Título** | Proceso de normalización de datos | | | |
| **Escena** | **Imagen** | **Sonido** | **Narración (voz en off)** | **Texto** |
| **01** | Imagen mostrando el proceso de normalización de datos.  Data report illustration concept  Fuente: <https://img.freepik.com/free-vector/data-report-illustration-concept_114360-883.jpg> | Música de fondo | El proceso de normalización de datos consiste en transformar los valores de la información de un formato no aceptado a un formato aprobado, que sea útil para su utilización en el proceso de ciencias de datos. | Normalización de datos. |
| **02** | Imagen que muestre proceso de eliminación de datos duplicados.  Stress concept illustration  Fuente: <https://img.freepik.com/free-vector/stress-concept-illustration_114360-2031.jpg> | Música de fondo | Al igual que en los procesos de limpieza de datos, se busca con la normalización eliminar la inexactitud e incoherencia de los conjuntos de datos, que se da, en muchos casos, debido a la integración de información de diversas fuentes. De la calidad y de la correcta realización de estas actividades dependerán en gran medida los resultados de los ejercicios de ciencia de datos que se realicen.  El proceso de normalización de datos se puede dividir en cuatro pasos: | Buscan eliminar la inexactitud y la incoherencia de los conjuntos de datos. |
| **04** | Imagen que muestre proceso que identifique las necesidades de la organización.  Data analysis concept illustration  Fuente: <https://img.freepik.com/free-vector/data-analysis-concept-illustration_114360-8053.jpg> | Música de fondo | **01. Definir la norma a aplicar:** el primer paso consiste en definir o identificar la norma que satisface las necesidades de la organización. La mejor manera de realizar esto es definir un modelo de datos para la organización, donde se representará el estado ideal al que deben ajustarse los datos de las diversas fuentes de información. En este modelo de datos se deberán identificar los activos de datos, como también definir los campos y sus detalles, los tipos de datos adecuados, los límites de caracteres, los patrones, los formatos (por ejemplo, los relacionados con las fechas), las unidades de medida y los dominios de valores (para los campos categóricos). | Definir la norma a aplicar: el primer paso consiste en definir o identificar la norma que satisface las necesidades de la organización. |
| **05** | Imagen que muestre proceso de verificación de datos.  Curiosity search concept illustration  Fuente: <https://img.freepik.com/free-vector/curiosity-search-concept-illustration_114360-11031.jpg> | Música de fondo | **02. Probar el estado de los datos:** una vez definida la norma a aplicar, se verifica que datos se ajustan a esta mediante el análisis de registros y atributos, la comparación y validación de patrones y el uso de diccionarios de datos. Este análisis debe finalizar con un informe sobre el perfil de la información, en el que se consolide cuántos de los datos analizados cumplen con las normas establecidas y cuáles deben transformarse. | Probar el estado de los datos: una vez definida la norma a aplicar, se verifica qué datos se ajustan a esta mediante el análisis de registros y atributos. |
| **06** | Imagen que muestre proceso de transformación de datos.  Text files concept illustration  Fuente: <https://img.freepik.com/free-vector/text-files-concept-illustration_114360-4402.jpg> | Música de fondo | **03. Transformar los datos:** el siguiente paso consiste en convertir los valores identificados como no conformes (que no cumplen las normas establecidas) en un formato ya normalizado. Esto puede incluir: transformación de los tipos de datos de los campos y de patrones y formatos (por ejemplo, los relacionados con correos electrónicos, que deben cumplir con un patrón definido, o el caso de las fechas, las cuales se deben manejar siempre en el mismo formato para que sean comparables), cambios de unidades de medida (como la conversión de los precios a una moneda establecida), ampliación de valores abreviados, eliminación de ruido presente en los valores de los datos (cadenas adicionales que no agregan valor) y reconstrucción de valores. Estas transformaciones se pueden realizar de forma manual o mediante herramientas automatizadas. | Transformar los datos: consiste en convertir los valores identificados como no conformes (que no cumplen las normas establecidas) en un formato ya normalizado. |
| **07** | Imagen que muestre prueba y validación de un proceso.  Product quality concept illustration  Fuente: <https://img.freepik.com/free-vector/product-quality-concept-illustration_114360-7301.jpg> | Música de fondo | **04. Probar nuevamente:** una vez aplicado el proceso de transformación para que los datos cumplan con la norma definida, se recomienda realizar nuevamente el proceso de prueba sobre el conjunto de datos, para detectar errores de estandarización. Igualmente, se aconseja utilizar los informes previos y compararlos con los últimos resultados para identificar en qué medida los errores de datos fueron corregidos en el proceso de transformación. | Probar nuevamente: se realiza una vez aplicado el proceso de transformación para que los datos cumplan con la norma definida. |
| **Nombre del archivo** | 228131\_V\_02 | | | |

## 3.2. Herramientas para transformación y limpieza de datos

|  |  |  |
| --- | --- | --- |
| **Tipo de recurso** | | Pestañas o tabs verticales |
| **Introducción** | | Al momento de realizar actividades de transformación y limpieza de datos, hay una gran variedad de herramientas para utilizar, algunas de código abierto y otras pagadas, que ayudarán a realizar este proceso de ciencias de datos, que, por lo general, ocupa gran parte del tiempo.  A continuación, de describen algunas de estas herramientas: |
| Ilustración vectorial de la limpieza de bases de datos. Pequeños desarrolladores con pantalla de monitor de limpieza de limpiaparabrisas grande.  Fuente: <https://www.shutterstock.com/es/image-vector/vector-illustration-database-cleansing-tiny-developers-1282907026>    **Imagen:** 228131\_i\_16 | | |
| **Python:** | este es uno de los lenguajes de programación más utilizados para proyectos de ciencias de datos. Se utilizan en la fase de limpieza y transformación de librerías, como NumPy, Pandas y Matplotlib, cada una con procedimientos y funciones de gran utilidad para manipular los conjuntos de datos. (<https://www.python.org/>) | |
| **NumPy:** | es una librería que se utiliza para realizar operaciones matemáticas requeridas en los conjuntos de datos. | |
| **Pandas:** | es una herramienta que se emplea para importar y gestionar los conjuntos de datos (*dataframes*). | |
| **Matplotlib:** | es una librería que ayuda a realizar gráficos a partir de los conjuntos de datos disponibles. | |
| **R:** | es un lenguaje de programación orientado especialmente a labores de análisis exploratorio de datos para tareas de ciencia de datos. El entorno de R proporciona un conjunto de librerías de gran utilidad para la manipulación de información, clasificación, generación de gráficos y técnicas estadísticas. Gracias a los paquetes de computación distribuida, R puede administrar grandes cantidades de datos en segundos y ejecutarse en diversos sistemas operativos. (<https://www.r-project.org/>) | |
| **OpenRefine:** | es una herramienta de código abierto para trabajar con datos desordenados. Realiza la manipulación de información mediante el uso de filtros y particiones en los conjuntos de datos a través del uso de expresiones regulares. (<https://openrefine.org/>) | |

|  |  |
| --- | --- |
| **Tipo de recurso** | Acordeón tipo 2 |
| **Introducción** | Otras herramientas importantes, en este caso de pago, son las siguientes: |
| La vista panorámica de múltiples bases de datos se coloca en tablas de bases de datos relacionales. Concepto de servidor de bases de datos, SQL, almacenamiento de datos, diseño de diagramas de bases de datos, Data center, Webhosting. Ilustración 3D.  Fuente: <https://www.shutterstock.com/es/image-illustration/panoramic-view-multiple-database-placed-on-2164135613>    **Imagen:** 228131\_i\_17 | |
| **Pentaho Data Integration:** es el componente de la plataforma Pentaho, orientado a las tareas de ETL. Entre otras funciones, comprende la migración de información entre aplicaciones o bases de datos, la exportación de información desde bases de datos o archivos planos, y herramientas de limpieza de datos (Hitachi, 2016). (<https://help.hitachivantara.com/Documentation/Pentaho/7.0/0D0/Pentaho_Data_Integration>) | |
| **Microsoft SQL Server Integration Services (SSIS):** esta herramienta de Microsoft proporciona sus propios servicios de integración de SQL Server para conectar con otros motores de bases de datos. Una de sus principales características es que permite una fácil migración a estructuras de datos, que ayuda a garantizar la consistencia de estos y disminuir los riesgos habituales de este tipo de operaciones (Microsoft, 2022). (<https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver16>) | |

# 4. Carga de datos

|  |
| --- |
| Cuadro de texto |
| En esta última fase de los procesos ETL, se cargan los datos recopilados y transformados en un formato ahora consistente y homogéneo. Estos datos se almacenan en un *data warehouse* (DW) o depósito de datos. Por lo general, el DWse encuentra separado de la base de datos de producción de la organización, lo que propende así a su uso y aprovechamiento sin afectar la labor diaria del negocio. El depósito de datos funciona como un repositorio central, con información que proviene de diversas fuentes de datos y de diferentes tipos (estructurados, semiestructurados y no estructurados). Una vez se encuentran en el depósito, los usuarios pueden acceder a estos mediante herramientas de inteligencia de negocios. |

## 4.1. *Data warehouse*: conceptos y herramientas

|  |
| --- |
| Cuadro de texto |
| Un *data warehouse*, o depósito de datos, es una plataforma utilizada para almacenar y procesar datos que provienen de diversas fuentes y formatos y que ocupa un lugar central en un sistema de inteligencia de negocios (DataScientest, 2022). Allí es donde se cargan los datos recolectados mediante las técnicas de extracción, después de que se transforman, para que tengan el formato y la calidad adecuados para su explotación en los procesos de ciencia de datos. Por lo general, estos depósitos de información se encuentran separados de las bases de datos de producción de la organización para no alterar la información original ni sobrecargar el sistema de información de producción.  Sumado a lo anterior, los depósitos de datos funcionan como repositorio central, y su información proviene de una o varias fuentes de datos, como son los sistemas de información de la organización, los sistemas transaccionales y las bases de datos relacionales. Al tener la información centralizada en un solo lugar, la organización logra una visión holística de sus sistemas y procesos claves para un mejor entendimiento y un proceso de toma de decisiones soportado en los datos. Además, los *data warehouse* están orientados al análisis de información histórica; por tanto, dependiendo de la antigüedad y tamaño de la organización, esta debe disponer de una gran capacidad de almacenamiento (en algunas organizaciones ya se trabaja en cantidades de *petabytes*), ya que muchos de los datos permanecen por largos periodos de tiempo. Por tal razón, cada vez es más frecuente el uso de soluciones en la nube para este proceso y evitar así que la organización deba realizar grandes inversiones en equipos e infraestructura para desarrollar sus procesos. Con el uso de la computación en la nube o infraestructura como servicio, las organizaciones solo pagan por los servicios usados y bajo demanda. |

|  |  |
| --- | --- |
| **Tipo de recurso** | Rutas / Pasos. Verticales 1 |
| **Introducción** | Continuando con una aproximación al concepto de *data warehouse* en esta tercera parte del proceso ETL, las siguientes son las características que debe tener un DW: |
| Sala de servidores isométrica, datos de almacenamiento en la nube, centro de datos, procesamiento de grandes datos y tecnología informática. Protección de datos, un escudo contra el fondo de la sala de servidores. Comunicación de la granja de servidores.  Fuente: <https://www.shutterstock.com/es/image-vector/server-room-isometric-cloud-storage-data-1794802207>  Nota:  **Imagen:** 228131\_i\_18 | |
| **Botón 1** | **Orientado a temas**: los datos contenidos en los DW están organizados por temáticas o entidades (cliente, vendedor, producto), que contienen únicamente información relevante para la toma de decisiones. |
| **Botón 2** | **Integrado:** las fuentes de datos de los DW son diversas, pero, al almacenarse en estos depósitos de datos, deben ser integradas y homogenizadas para que puedan compararse y utilizarse en conjunto. |
| **Botón 3** | **Variable con el tiempo**: los DW mantienen los datos históricos, a diferencia de los sistemas de información transaccionales, que solo almacenan la información reciente. En estos depósitos, los datos pueden permanecer por años según las necesidades y políticas de la organización. |
| **Botón 4** | **No volátil**: los usuarios no deben cambiar ni actualizar los datos almacenados en los DW. En caso de ser necesario, los cambios en los datos solo podrán realizarlos los profesionales TI de la organización. |
| **Botón 5** | **Multidimensional**: los DW pueden utilizar una estructura relacional o multidimensional para almacenar los datos. Una representación típica de estas estructuras son los cubos de datos, donde las dimensiones del negocio o los temas son aristas en el cubo de datos. |
| **Botón 6** | **Metadatos**: los DW contienen metadatos (datos acerca de los datos), que facilitan la gestión de los datos complejos allí almacenados. |

|  |  |
| --- | --- |
| **Tipo de recurso** | Infografía estática |
| **Texto introductorio** | Existen otro tipo de depósitos de datos conocidos como *data lakes* o lago de datos. Estos son repositorios de información que se encuentran almacenados en su forma nativa, es decir, sin procesar (estructurados, semiestructurados o no estructurados). Estos almacenan no solo la información que se utiliza en el momento, sino también los datos históricos.  La siguiente infografía presenta las principales diferencias entre los *data lakes* y los *data warehouse*: |
| Gráfico, Gráfico de embudo  Descripción generada automáticamente  Fuente: <https://www.7puentes.com/blog/wp-content/uploads/2021/08/Post_42_2-630x1024.jpg> | |
| **Código de la imagen** | 228131\_i\_19 |

## 4.2. Tecnologías de almacenamiento (OLAP, ROLAP, MOLAP, HOLAP)

|  |  |
| --- | --- |
| Tipo de recurso | Cajón de texto de color |
| OLAP es el acrónimo en inglés de procesamiento analítico en línea (*OnLine Analytical Processing*) y su objetivo es agilizar las consultas de grandes cantidades de datos (Muñoz, 2018). Para ello, utiliza estructuras multidimensionales, también conocidos como cubos OLAP, que almacenan información de grandes bases de datos. Sus principales usos son en procesos de minería de datos, ventas, *marketing* e inteligencia de negocios. El creador del concepto OLAP fue Edgar Frank Codd, conocido también por ser precursor de las bases de datos relacionales. Su propuesta inicial consistió en tener una disposición de los datos en vectores para permitir un análisis rápido, (Joyanes, 2019).  Los cubos OLAP se suelen comparar con una hoja de datos ampliada. Este cubo es una base de datos que tiene varias dimensiones, que amplían las posibilidades que ofrecen en la actualidad las hojas de cálculo de solo dos dimensiones. Los OLAP pueden a su vez contener varios cubos o vectores, que también aumentan las posibilidades del sistema (estos son conocidos como hipercubos). Por lo general, los cubos OLAP se componen de tres dimensiones, pero pueden albergar más si es necesario, por el tipo de organización o por la información almacenada.    Fuente: <https://geeks.ms/lmblanco/wp-content/uploads/sites/117/2016/05/CubosDatosSQLServer2008_01.jpg>    **Imagen:** 228131\_i\_20 | |

|  |  |  |
| --- | --- | --- |
| **Tipo de recurso** | Tarjetas Avatar | |
| **Introducción** | En el procesamiento analítico en línea también hay disponibles variadas herramientas. Algunas de ellas se caracterizan por | |
| Ofrecer una visión multidimensional de los datos. | |  |
| No imponer restricciones con relación al número de dimensiones a manejar. | |  |
| Permitir la definición de forma flexible sobre las dimensiones: restricciones, agregaciones y jerarquías entre ellas. | |  |
| Ofrecer operadores intuitivos de manipulación. | |  |
| Permitir consultas a la herramienta OLAP, seleccionando atributos del esquema multidimensional. | |  |

|  |  |  |
| --- | --- | --- |
| **Tipo de recurso** | Rutas /Pasos Horizontal | |
| **Introducción** | Las herramientas OLAP se clasifican en tres grupos: MOLAP (procesamiento analítico multidimensional en línea), ROLAP (procesamiento analítico en línea relacional) y HOLAP (procesamiento analítico en línea híbrido). | |
| **Paso 1** | **MOLAP:** la información es precalculada y luego almacenada en cubos de datos multidimensionales. Los datos se organizan en una estructura tipo cubo que el usuario puede rotar. Es el más usado para resúmenes e informes financieros. | Proceso OLAP. Concepto de procesamiento analítico en línea, análisis de datos, servidor OLAP, BI de Business Intelligence, KPI de indicadores clave de rendimiento, panel, SQL, MDX, representación 3D.  Fuente: <https://www.shutterstock.com/es/image-illustration/olap-process-online-analytical-processing-concept-2187453469>  Nota:  **Imagen:** 228131\_i\_21 |
| **Paso 2** | **ROLAP:** los datos son introducidos directamente desde el *data warehouse* u otra fuente de datos relacional y no son almacenados por separado. La base de datos relacional maneja los requerimientos de DW, y el motor de ROLAP proporciona la función analítica. |  |
| **Paso 3** | **HOLAP:** trata de combinar las ventajas de las soluciones MOLAP y ROLAP. Almacena algunos datos de un motor relacional, mientras que otros los compendia en una base de datos multidimensional. |  |

## 4.3. Mantenimiento del*data warehouse*

|  |
| --- |
| Cuadro de texto |
| La fase de mantenimiento del DWes un proceso que implica el monitoreo, evaluación, reparación y mejora del sistema. Por lo general, los sistemas de información y, en este caso, los depósitos de datos necesitan actualizarse con nuevos productos, estándares, normas, regulaciones o reglas del negocio. En este sentido, es fundamental que las herramientas seleccionadas sean flexibles y provean las funciones necesarias para realizar estos procesos. Los procesos de mantenimiento de DW, como pueden ser actualización de las vistas o ingreso de nuevos datos, deben ser realizados por profesionales TI con experiencia en este tipo de actividades, porque una modificación no planificada o que se realice de forma incorrecta puede acarrear pérdida de información. |

|  |  |  |
| --- | --- | --- |
| Tipo de recurso | Rutas / Pasos. Verticales 2 | |
| **Introducción** | Durante todo el ciclo de vida del proceso de ciencia de datos, es importante realizar procesos de evaluación sobre el sistema. Esta evaluación debe ser periódica y sistemática para determinar si el sistema está funcionando de la forma esperada y con eficiencia y si se está cumpliendo con los requerimientos y necesidades de la organización. La evaluación también permitirá identificar necesidades de actualización para dar respuesta a las necesidades cambiantes del proceso.  Existen tres tipos de mantenimiento para considerar: depuración, actualización y adición de nuevas funcionalidades. | |
| **1** | **Depuración:** consiste en la detección, prueba y reparación de errores. Muchas de las fallas son corregidas durante el proceso de implementación, pero algunas solo se detectan en fase de producción. Así que es necesario realizar los ajustes requeridos para que el *data warehouse* funcione de la forma esperada. | Tecnología de desarrollo de aplicaciones, depuración y prueba de código de software para errores y fallos de búsqueda. Concepto de datos de programación, análisis y codificación. Ilustración vectorial.  Fuente: <https://www.shutterstock.com/es/image-vector/technology-application-development-debugging-code-testing-1868752312>  Nota:  **Imagen:** 228131\_i\_22 |
| **2** | **Actualización:** las condiciones y necesidades de la organización cambian con el tiempo. Por ende, es necesario que el *data warehouse*, sus vistas y la información allí almacenada se adapten a estos cambios para que sean de utilidad. | Ilustración del concepto de actualización del sistema operativo  Fuente: <https://www.freepik.es/vector-gratis/ilustracion-concepto-actualizacion-sistema-operativo_13416255.htm#query=actualizaci%C3%B3n&position=3&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_23 |
| **3** | **Adición:** en algunos casos puede ser necesario agregar nuevas funcionalidades o vistas en el *data warehouse*, sin afectar su funcionamiento. | Concepto de ciudad inteligente digital, transmisión y procesamiento de big data, almacén de centro de datos  Fuente: <https://www.freepik.es/vector-gratis/concepto-ciudad-inteligente-digital-transmision-procesamiento-big-data-almacen-centro-datos_3629619.htm#query=vistas%20data%20warehouse&position=0&from_view=search&track=sph>  Nota:  **Imagen:** 228131\_i\_24 |

**Síntesis**

|  |  |
| --- | --- |
| **Tipo de recurso** | Síntesis |
| Gestión analítica de datos  Síntesis: procesos ETL (extracción, transformación y carga) | |
| **Introducción** | En la siguiente figura se encuentran a manera de síntesis todos los temas vistos y aprendidos en este componente formativo, lo que servirá como guía para recordar cada contenido y así realizar de manera satisfactoria las respectivas actividades a desarrollar. |
| **Figura 3**  *Síntesis de procesos ETL (extracción, transformación y carga)*  Diagrama  Descripción generada automáticamente | |

**Actividad interactiva**

|  |  |  |
| --- | --- | --- |
| **Tipo de recurso** | Actividad didáctica. Arrastrar y soltar | |
| Apreciado aprendiz, a continuación, encontrará una actividad didáctica para afianzar los conceptos y aprendizajes del componente de formación 15: “Procesos ETL (extracción, transformación y carga)”. Lea con cuidado cada definición y arrástrela al concepto correcto.  **Retroalimentación respuesta positiva:** ¡Felicitaciones! Ha logrado completar con éxito las frases de esta actividad, lo que identifica una apropiación de conocimiento de lo expuesto en este módulo académico.  **Retroalimentación respuesta negativa:** ¡Lo sentimos! No ha logrado completar las frases de manera satisfactoria. Por favor, revise nuevamente el material formativo e inténtelo de nuevo. | | Hand drawn speech therapy illustration  Fuente: <https://img.freepik.com/free-vector/hand-drawn-speech-therapy-illustration_23-2149210238.jpg>  Nota:  **Imagen:** 228131\_i\_25 |
| Son una parte esencial de los procesos de ciencia de datos, donde se aplican diferentes herramientas, técnicas y actividades que permiten extraer, transformar y cargar información. | | Procesos ETL |
| Consiste en recopilar desde las diversas fuentes de información los datos requeridos para la realización del proyecto de ciencias de datos. Las fuentes de información son, en la mayoría de casos, muy diversas, e incluyen archivos físicos, como facturas, registros en papel, libros contables y respaldo, o también digitales, como hojas de cálculos, documentos de texto, base de datos, entre otros. | | Extracción de datos |
| Consiste en desarrollar algoritmos o usar herramientas que permitan la recuperación de documentos e información de una página web de forma automatizada. | | *Scraping* o *web scraping* |
| Permite la conversión electrónica o mecánica de imágenes de texto mecanografiado, manuscrito o impreso en texto codificado por la máquina, ya sea a partir de un documento escaneado, de una foto de un documento, de una foto de escena o de un texto de subtítulos superpuesto a una imagen. | | Reconocimiento óptico de caracteres |
| Es un proceso que utiliza técnicas matemáticas, estadísticas y de aprendizaje de máquina para identificar y extraer información útil de las enormes cantidades de datos existentes, para luego convertir dicha información en conocimiento en forma de patrones. | | Minería de datos |
| En esta fase se realizan actividades de validación, eliminación de duplicados, codificación, ajustes de formato y demás requeridas, para así identificar qué datos son útiles y cuáles son deficientes para el proceso a iniciar. | | Transformación de datos |
| Es el proceso de transformar los valores de los datos de un formato no aceptado a un formato aceptado y apto para su uso. | | Normalización de datos |
| Es una plataforma que se utiliza para almacenar y procesar datos provenientes de diversas fuentes y formatos y ocupa un lugar central dentro de un sistema de inteligencia de negocios. | | *Data warehouse* |
| Su objetivo es agilizar las consultas de grandes cantidades de datos, para lo cual utiliza estructuras multidimensionales. | | OLAP |
| Conjunto de preguntas preparadas sobre aspectos de interés para el proyecto de ciencia de datos a realizar. Este puede estar dirigido a toda la organización o a una muestra de esta. | | Cuestionario |

# Material complementario

|  |  |  |  |
| --- | --- | --- | --- |
| Tipo de recurso | Material complementario | | |
| Tema | Referencia APA del material | Tipo | Enlace |
| Herramientas de extracción de datos | Digital Methods Initiative. (2020). *DMI Tools*. | Sitio web | <https://wiki.digitalmethods.net/Dmi/ToolDatabase> |
| ABC del OLAP | Evaluando Software. (2022). *ABC del OLAP*. | Sitio web | <https://www.evaluandosoftware.com/abc-del-olap/> |
| Ciencia de datos | García, J., Molina, J. M., Berlanga, A., Patricio, M. A., Bustamante, A. L. y Padilla, W. R. (2018). *Ciencia de datos: técnicas analíticas y aprendizaje estadístico en un enfoque práctico*. Altaria. | Libro | <https://www.amazon.com/-/es/Jes%C3%BAs-Garc%C3%ADa-Herrero/dp/8494731963> |
| CRISP-DM | IBM. (2021). *Conceptos básicos de ayuda de CRISP-DM*. | Sitio web | <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview> |
| Ciencia de datos | Jones, H. (2019). *Ciencia de los datos: La guía definitiva sobre análisis de datos, minería de datos, almacenamiento de datos, visualización de datos, big data para empresas y aprendizaje automático para principiantes*. | Libro | <https://www.amazon.com/-/es/Herbert-Jones/dp/109080296X> |
| Ciencia de datos | Martín, I. y Fernández, A. (2020). *Ciencia de datos para la ciberseguridad.* Edición Ra-Ma. | Libro | <https://bit.ly/3U2jvqQ> |
| Procesos ETL | Microsoft. (s. f.). *Extracción, transformación y carga de datos (ETL)*. | Sitio web | <https://learn.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl> |
| Bases de datos | Orbegozo, B. (2013). *Gestión de bases de datos con SQL, MySQL y Access: curso práctico*. Alfaomega; Altaria. | Libro | <https://bit.ly/3gsxAjD> |
| Gestión de las TSI | Piattini, M. y Ruiz, F. (2020). *Gobierno y Gestión de las Tecnologías y los Sistemas de Información*. Ra-Ma. | Libro | <https://www.ra-ma.es/libro/gobierno-y-gestion-de-las-tecnologias-y-los-sistemas-de-informacion_104970/> |

# Glosario

|  |  |
| --- | --- |
| **Tipo de recurso** | Glosario |
| CRISP-DM: | la *Cross Industry Standard Process for Data Mining* es una metodología estandarizada para la implementación de proyectos de ciencias de datos y minería de datos. |
| *Data lake*: | “es un repositorio centralizado que le permite almacenar todos sus datos estructurados y no estructurados a cualquier escala. Puede almacenar sus datos sin modificarlos y sin tener que estructurarlos primero” (AWS, s. f., párr. 1). |
| *Data warehouse*: | también conocido como depósito de datos, es un repositorio unificado para todos los datos que recogen los diversos sistemas de una organización. |
| ETL: | *extract, transform and load* (extraer, transformar y cargar). |
| Inteligencia de negocios: | la *business intelligence* (BI) “combina análisis de negocios, minería, visualización, herramientas e infraestructura de datos, además de prácticas recomendadas para ayudar a las empresas a tomar decisiones basadas en los datos” (Tableau, s. f., párr. 1) |
| JSON: | “es un formato de archivo estándar abierto y un formato de intercambio de datos que utiliza texto legible para almacenar y transmitir objetos de datos que constan de pares atributo-valor y tipos de datos de matriz” (Ministerio de Ambiente y Desarrollo Sostenible, s. f., definición). |
| *Metadata*: | son los datos que proporcionan información sobre uno o más aspectos de los mismos. Se utilizan para resumir información básica de estos, que puede facilitar el seguimiento y el trabajo con datos específicos. |
| OCR: | en español, reconocimiento óptico de caracteres, es una tecnología que reconoce el texto dentro de una imagen digital. Se utiliza habitualmente para reconocer texto en documentos e imágenes escaneadas. |
| OLAP: | en español, procesamiento analítico en línea. Es una solución empleada en el campo de la inteligencia de negocios, cuyo objetivo es agilizar la consulta de grandes cantidades de datos (Muñoz, 2018). |
| *Scraping*: | se refiere al proceso de extracción de contenidos y datos de sitios web mediante *software* o algoritmos. |

# Referencias bibliográficas

|  |  |
| --- | --- |
| **Tipo de recurso** | Bibliografía |
| AWS. (s. f.). *¿Qué es un lago de datos?* <https://go.aws/3AH2U4X> | |
| DataScientest. (2022). *Data Warehouse: ¿qué es y cómo utilizarlo?* <https://bit.ly/3XwFXLo> | |
| Hitachi. (2016). *Pentaho Data Integration.* <https://htchivantara.is/3AGs8jX> | |
| Joyanes, L. (2019). *Inteligencia de negocios y analítica de datos. Una visión global de business intelligence & analytics*. Alfaomega. | |
| Microsoft. (2022). *SQL Server Integration Services*. <https://bit.ly/3EWjD6K> | |
| Ministerio de Ambiente y Desarrollo Sostenible. (s. f.). *JavaScript Object Notation (JSON)*. <https://bit.ly/3EUhs2D> | |
| Muñoz, A. (2018). *OLAP. Resumen final.* Centro de Producción de Contenidos Digitales.<https://repositorio.upct.es/handle/10317/7168> | |
| Rebollo, M. (2017). *Minería de datos. Usando la tecnología a nuestro favor*. <https://www.gestiopolis.com/mineria-datos-usando-la-tecnologia-favor/> | |
| Tableau. (s. f.). *¿Qué es la inteligencia de negocios? Guía sobre la inteligencia de negocios y por qué es importante*. <https://www.tableau.com/es-es/learn/articles/business-intelligence> | |