**FORMATO PARA EL DESARROLLO DE COMPONENTE FORMATIVO**

|  |  |
| --- | --- |
| PROGRAMA DE FORMACIÓN | Desarrollo e implementación de soluciones para la transformación digital |

|  |  |  |  |
| --- | --- | --- | --- |
| COMPETENCIA | 220501114- Sistematizar datos masivos de acuerdo con métodos de analítica y herramientas tecnológicas. | RESULTADOS DE APRENDIZAJE | 220501114-02 - Integrar datos según técnicas de tratamiento de datos. |

|  |  |
| --- | --- |
| NÚMERO DEL COMPONENTE FORMATIVO | CF14 |
| NOMBRE DEL COMPONENTE FORMATIVO | Proceso de integración de datos y ETL |
| BREVE DESCRIPCIÓN | El recurso educativo presenta los conceptos, teorías, técnicas y herramientas empleadas en sistematización de datos. Se dan las pautas de las metodologías y paradigmas usados para el proceso ETL, el cual quizás, es el más importante a nivel técnico en la gestión de información, para la transformación y carga que deben realizarse para la visualización y análisis de datos. |
| PALABRAS CLAVE | Analítica, datos, ETL, inteligencia de negocio, origen de datos |

|  |  |
| --- | --- |
| ÁREA OCUPACIONAL | Ventas y servicios |
| IDIOMA | Español |

1. **TABLA DE CONTENIDOS:**

**Introducción**

**1. Extracción y minería de datos**

**2. Técnicas de limpieza y transformación de datos**

2.1 Open Refine

2.2 Astera

**3. Modelo de datos transaccionales**

3.1 OLTP

3.2 SQL

3.3 NO-SQL

**4. Bodegas de datos**

4.1 Diseño de mercados

4.2 Hechos, dimensiones

4.3 Cubos OLAP y ROLAP

4.4 Llenado de almacén de datos

**5. Herramientas tecnológicas de ETL**

1. **INTRODUCCIÓN**

En este componente se abordarán los conceptos y fundamentos del proceso de extracción, transformación y carga de datos, con el fin de procesar la información para la visualización y análisis, que permita la toma de decisiones. A continuación, se presenta un video que contextualiza al respecto:

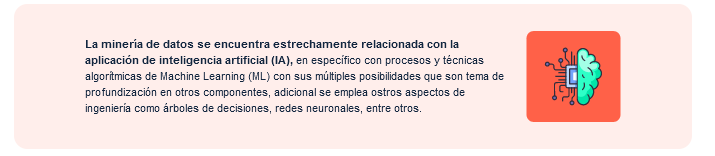
Agregar video

1. **DESARROLLO DE CONTENIDOS:**

## 1. Extracción y minería de datos

La identificación de las fuentes de datos es el comienzo técnico para la extracción de todos los datos necesarios para los reportes y la creación de los tableros de mando, permitiendo visualizar la información de la organización. Su contraparte, es decir, el final de muchos ciclos de inteligencia de negocios, es la minería de datos, la cual se presenta como una de las alternativas luego de la extracción y limpieza de datos.

Los datos encierran patrones y comportamientos de los que es posible extraer conocimiento sobre los eventos que los han generado. Frecuentemente, las cosas no son como aparentan ser. Por esta razón, la visualización de datos, aunque necesaria, no es del todo suficiente para llegar hasta el conocimiento que se esconde detrás de estructuras y relaciones poco superficiales en los datos (Gorenés, Casas, & Minguillón, 2017).



Este proceso consiste en tomar o extraer información requerida por las entidades de diferentes bases de datos o fuentes, su nombre está asociado a la minería ya que la extracción es vital en ambas actividades; por otro lado, su propósito es extraer datos útiles que ayuden a la resolución de las diferentes situaciones que se originan en las entidades y su gestión se puede hacer a través de herramientas estadísticas tanto básicas como avanzadas que proporciona la inteligencia artificial.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Interfaz de usuario gráfica

Descripción generada automáticamente

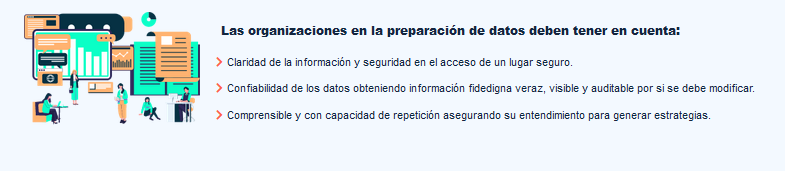
Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

## 2. Técnicas de limpieza y transformación de datos

Una de las inversiones más largas, a menudo, suele ser el proceso de limpieza de datos; sin duda es necesario realizar por cada fuente de datos una revisión sobre los datos, tales como características, bases de datos, tamaño, formato de cada registro, unicidad según cada campo, completitud, consistencia entre tablas, etc. Determinar calidad de los datos y verificar qué elementos requieren limpieza. Es normal que en cada fuente de datos se deba aplicar un proceso de limpieza y refinación de los datos diferente.

El proceso de limpieza de datos es muy importante para contar con los datos adecuados que van a ser utilizados, consultados, investigados, extraídos o buscados con el propósito de que la información sea precisa y válida para los diferentes análisis que se requieren hacer en el proceso, garantizando la seguridad de los datos obtenidos.



Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

### 2.1 Open Refine

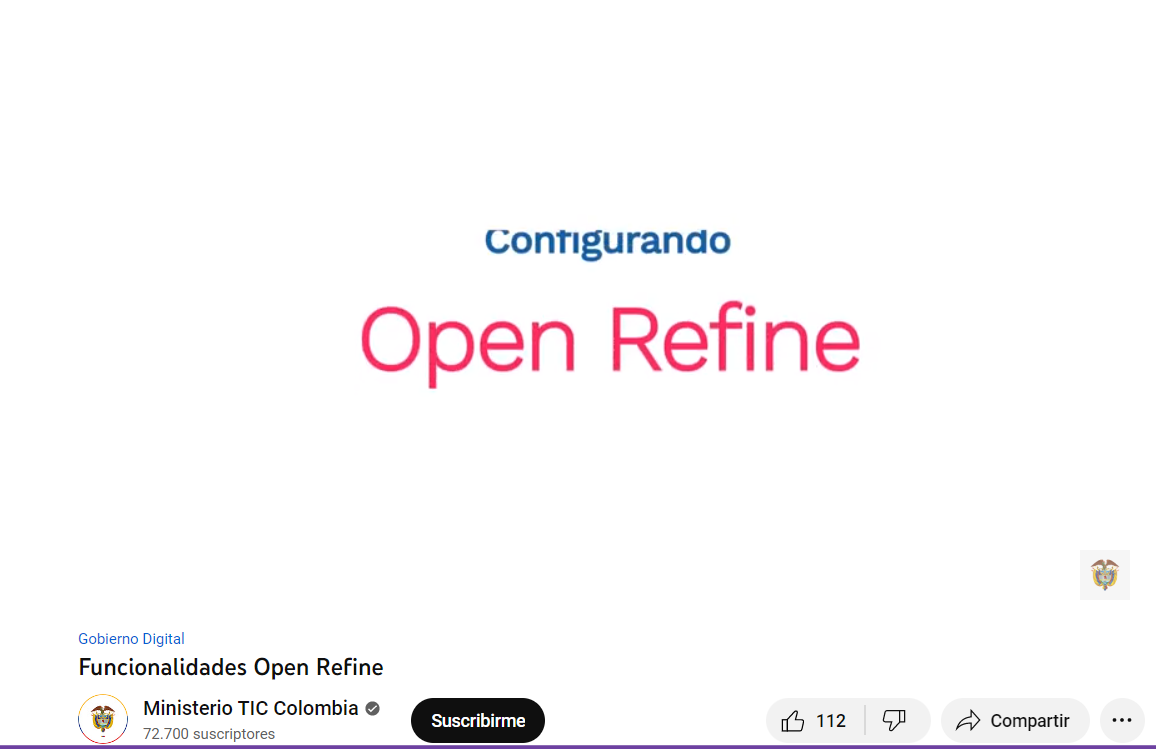
Es una aplicación de descarga libre y código abierto muy útil para realizar estas tareas de limpieza de datos y adaptarlos a los formatos y condiciones previas a otros procesos propios de ETL o minería de datos.

Interfaz de usuario gráfica, Texto, Sitio web

Descripción generada automáticamente

(<https://openrefine.org/>)

Como complemento, puede visitar el siguiente enlace:

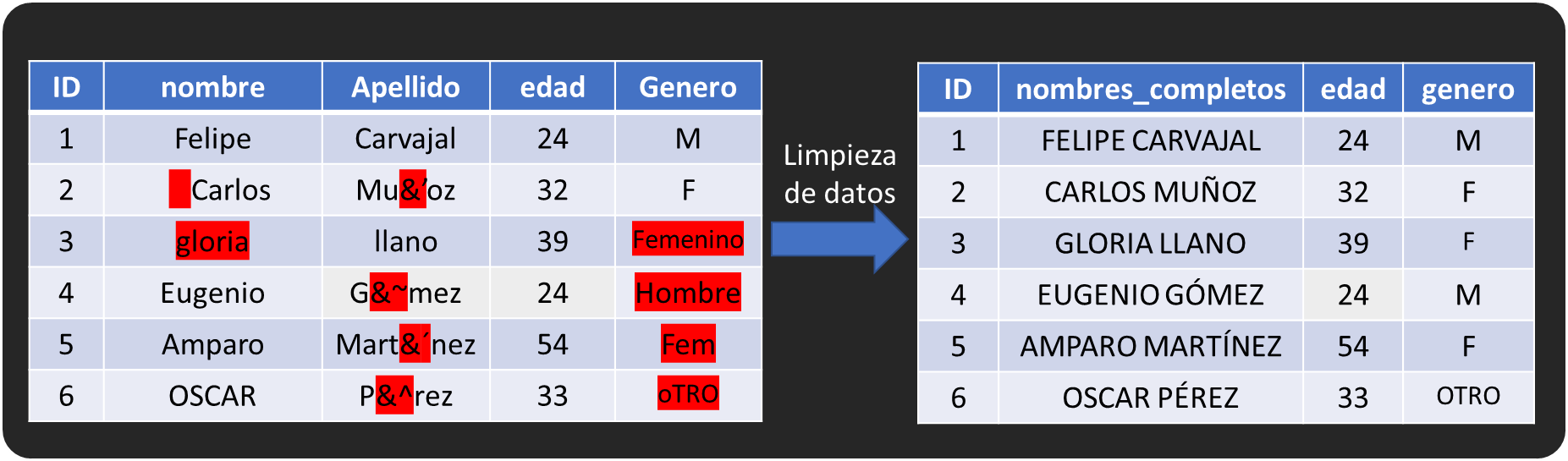


Su uso es relativamente fácil, lo importante es tener claridad de los conceptos y funcionalidades que se requieren. Para la transformación y refinamiento de datos usa su propio lenguaje (*General Refine Expression Language -GREL*), sin embargo, como ventaja, se puede seleccionar entre otras opciones de lenguajes para codificar las expresiones de transformación de datos.

La aplicación permite adicionar columnas basadas en expresiones, dividir campos, ordenar, quitar columnas, reorganizarlas, renombrarlas y realizar otras transformaciones.

**Figura 1**

*Limpieza y conciliación de datos*

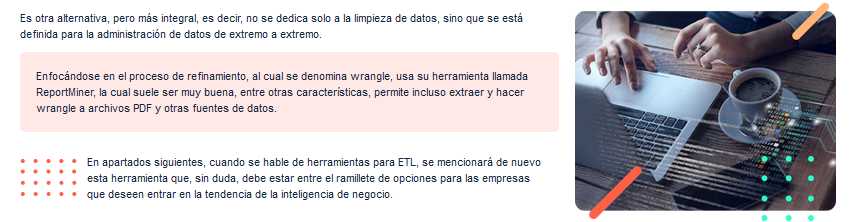


En el ejemplo anterior, se refinan los datos de manera que la información en los sistemas analíticos tenga una homogeneidad de formatos y estructuras, independiente de la fuente transaccional o archivos de entrada.

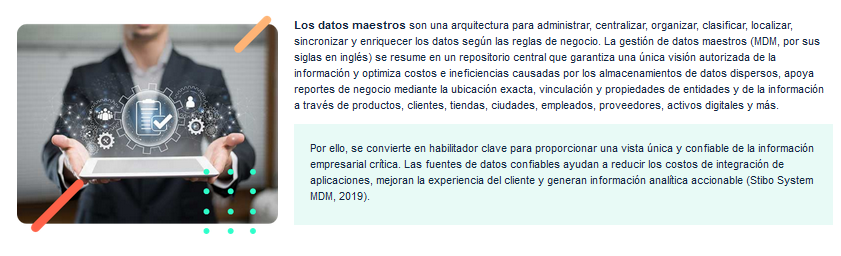
Para valores numéricos también se presentan múltiples funcionalidades, entre otras, la aplicación muestra valores poco típicos entre la colección de las columnas, ejemplo, si se tiene en fecha de nacimiento un año que marque por ejemplo 1890 (lo más probable es que se quería digitar 1990), sería una fecha atípica y se presenta una desviación muy grande en relación con el promedio de fechas registradas en la gran mayoría de registros.

### 2.2 Astera

Es otra alternativa, pero más integral, es decir, no se dedica solo a la limpieza de datos, sino que se está definida para la administración de datos de extremo a extremo.



Una de las maneras para garantizar que los datos tengan integridad y unicidad entre varios sistemas transaccionales y el sistema analítico mismo es el empleo de datos maestros (***Master Data Management***).



Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

**Figura 2**

*Ventajas de implementar Datos Maestros en los sistemas de información*



Se deben tener presentes estas ventajas al momento de implementar la gestión de los datos maestros.

## 3. Modelo de datos transaccionales

****

****

### 3.1 OLTP

Interfaz de usuario gráfica, Texto, Chat o mensaje de texto

Descripción generada automáticamente

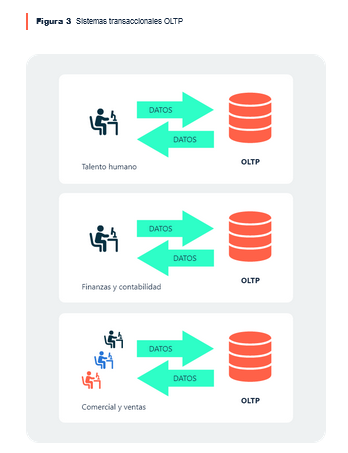
Otro tipo de aplicaciones de apoyo, son aquellas que se emplean en los procesos de producción, por ejemplo, si una compañía se dedica a la comercialización de productos de infraestructuras y equipos eléctricos de gran envergadura, tendrá un área de mercadeo, otra de ofertas, licitaciones y ventas.

Texto

Descripción generada automáticamente

Adicionalmente, se debe llevar registro no solo de quienes cierran el negocio, es clave en este caso, tener todos los datos entre la campaña *marketing*, eficiencia de las ofertas y eficacia de los vendedores para determinar el porcentaje de éxito del negocio.

Los datos que se surgen de estos sistemas serán claves para el proceso de centralización, limpieza e integración de datos del negocio y como ejemplo se expone en la siguiente figura.



Estos son ejemplos de sistemas transaccionales basados en OLTP que, generalmente, existen en las organizaciones.

También hay otras fuentes de información que se deben tener en cuenta, pues no todas las áreas funcionan con aplicaciones conectadas a bases de datos, además, existen procesos que podrían llegar a ser muy eficientes y con información adecuada que alimente el sistema de inteligencia de negocios con el uso de otros recursos, esto se conoce como fuentes de no transaccionales y las más comunes son:



### Una captura de pantalla de un videojuego Descripción generada automáticamente con confianza media3.2 SQL

Lenguaje estructurado de datos (*Structured Query Language*); es un lenguaje de programación para la gestión de bases de datos, a través de sentencias SQL, se accede y manipula datos de cualquier base de datos relacional del mercado, entre las más comunes MySQL, ORACLE, DB2, SQL SERVER, etc.

Como todo lenguaje, se compone de sentencias, cada una con utilidades y funciones diferentes; para administrar bases de datos y ejecutar sentencias SQL es necesario también un ambiente de desarrollo llamado IDE; existen múltiples opciones, las más recomendadas para descargar son:

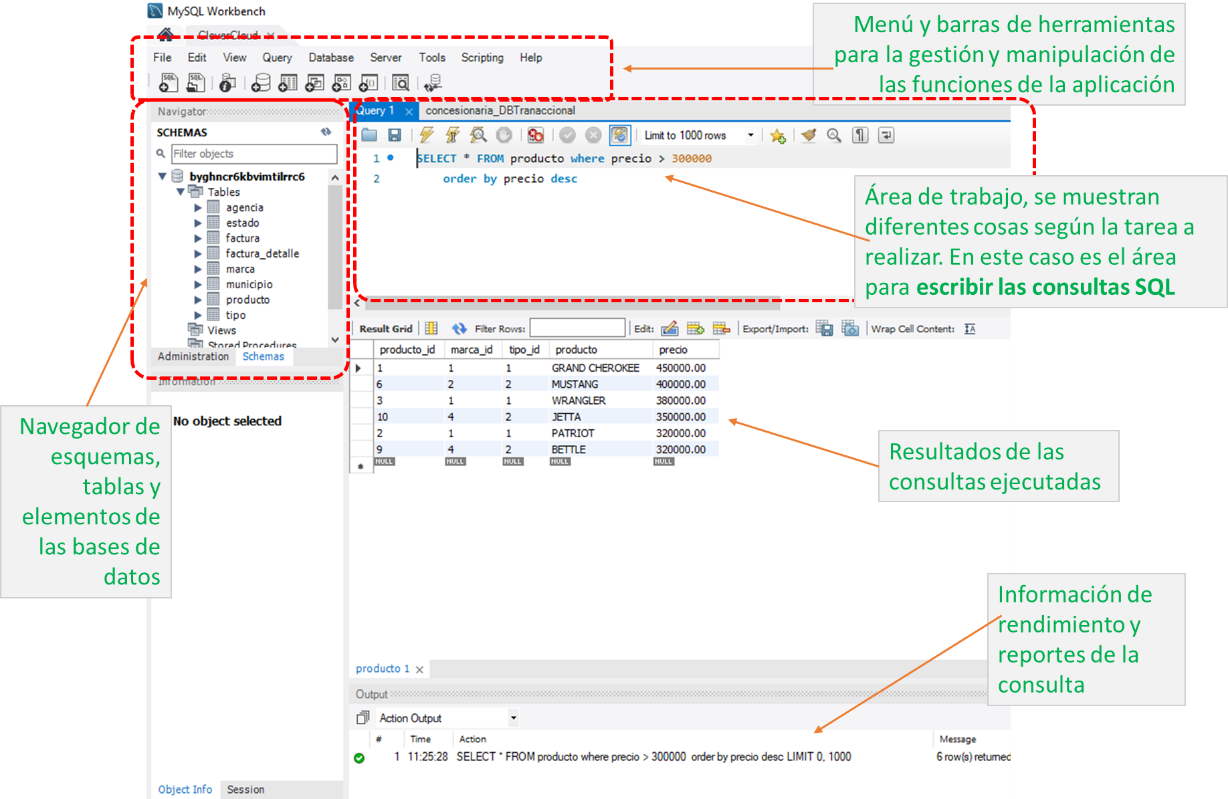
Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Independientemente del ambiente y herramienta seleccionada (ya sea de uso libre o de pago), la mayoría cuentan con el siguiente esquema:

**Figura 4.**

*Ambiente de las herramientas de gestión de bases de datos.*



A continuación, se presentan **las sentencias de gestión de estructura de datos,** más usadas**:**

****

Para llevar a la práctica los primeros pasos con las sentencias SQL, se usará un recurso en línea que, además de tener el ambiente gráfico para construir y ejecutar sentencias SQL, ya tiene una base de datos precargada para manipular los datos.



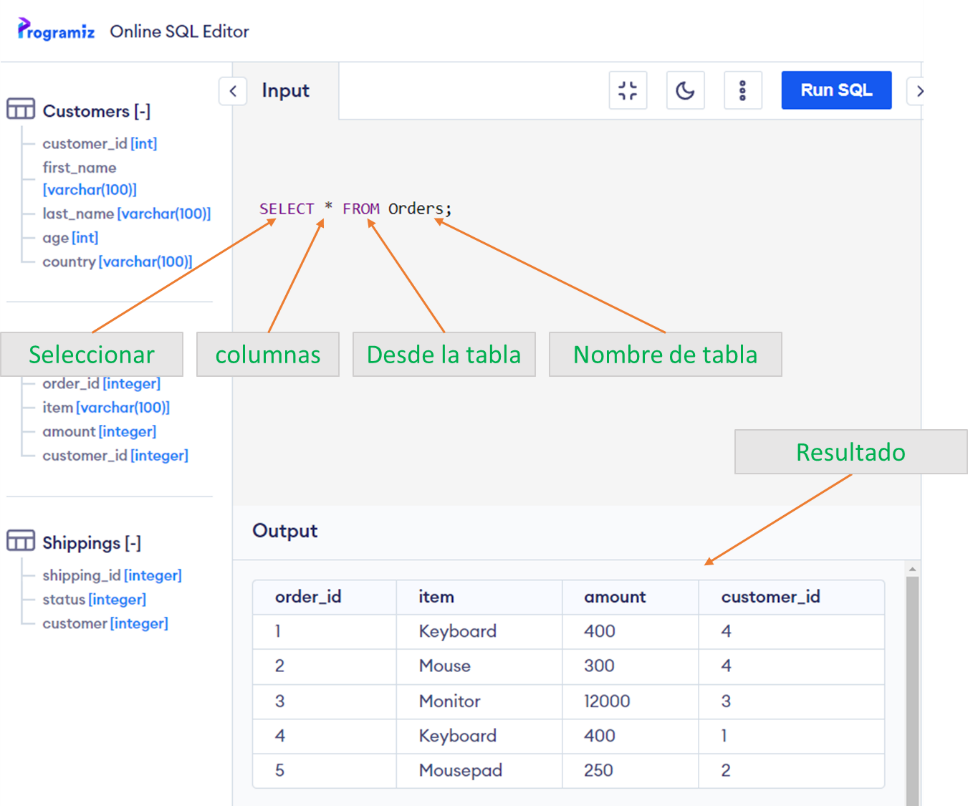
Este ambiente cuenta con una base de datos que contiene tres tablas: *customers, orders* y *shipings*. Realiza los siguiente ejercicios y práctica de las sentencias SQL que se proponen a continuación:

Texto

Descripción generada automáticamente con confianza media

**Figura 5**

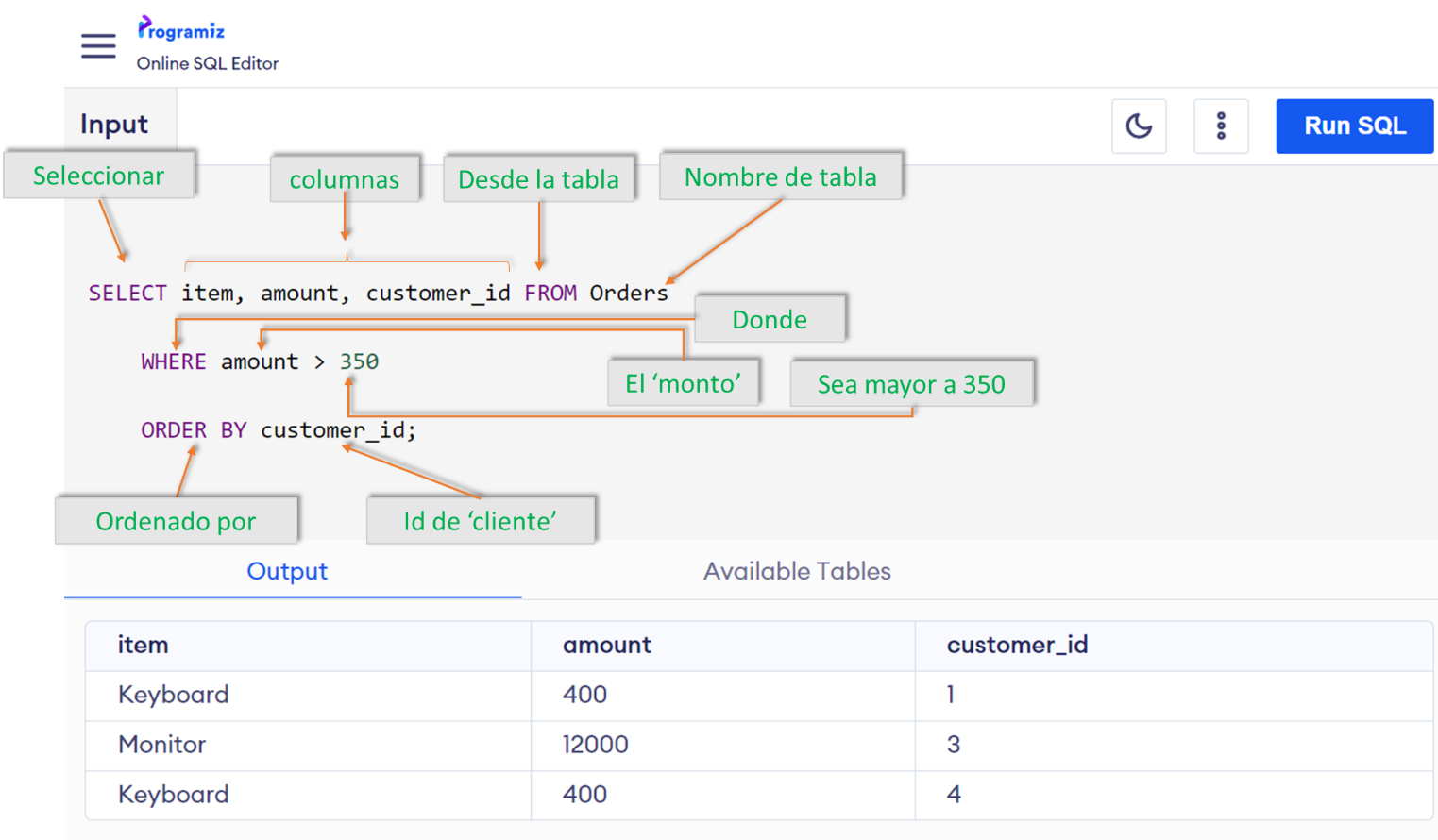
*Sentencia simple de SQL*



Ahora, se ejecuta una sentencia que solo llame ciertas columnas de la tabla, con una condición: que filtre los resultados y que organice los datos a razón de una de las columnas.

**Figura 6**

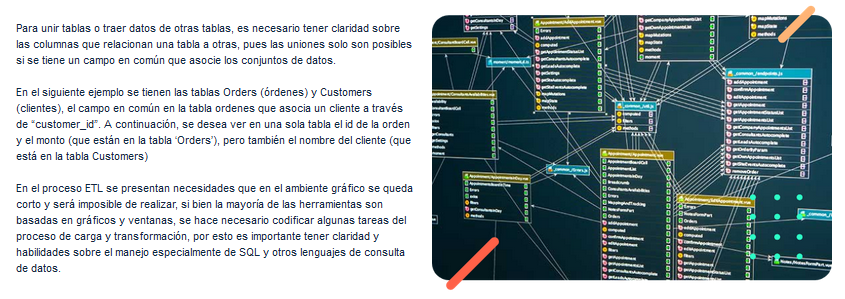
*Sentencia SQL con condición y ordenamiento*



Para agregar un nuevo registro, debe usar la sentencia: INSERT INTO.

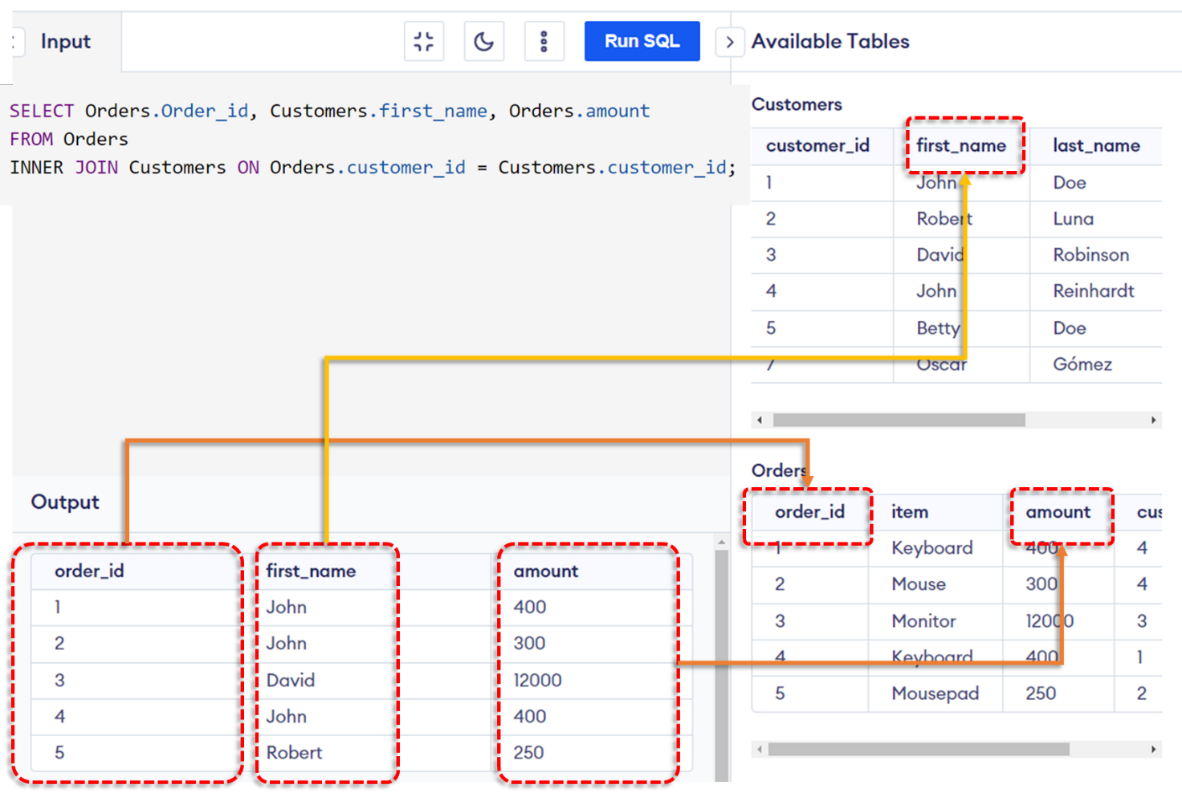
**Figura 7**

*Insertar registro en SQL*

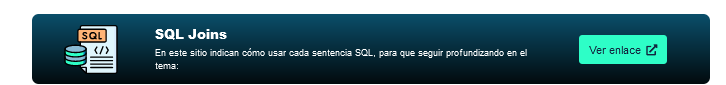


**Figura 8**

*JOIN - Uniones en SQL*



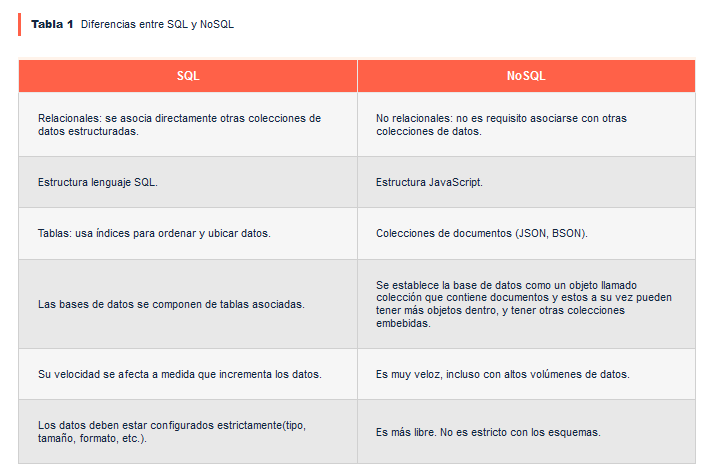
Ahora, se invita a realizar sus propias uniones y practicar diferentes sentencias SQL.



Para dominar este lenguaje, existen múltiples recursos en línea al respecto y también puede consultar en material complementario recomendado.

### 3.3 No-SQL

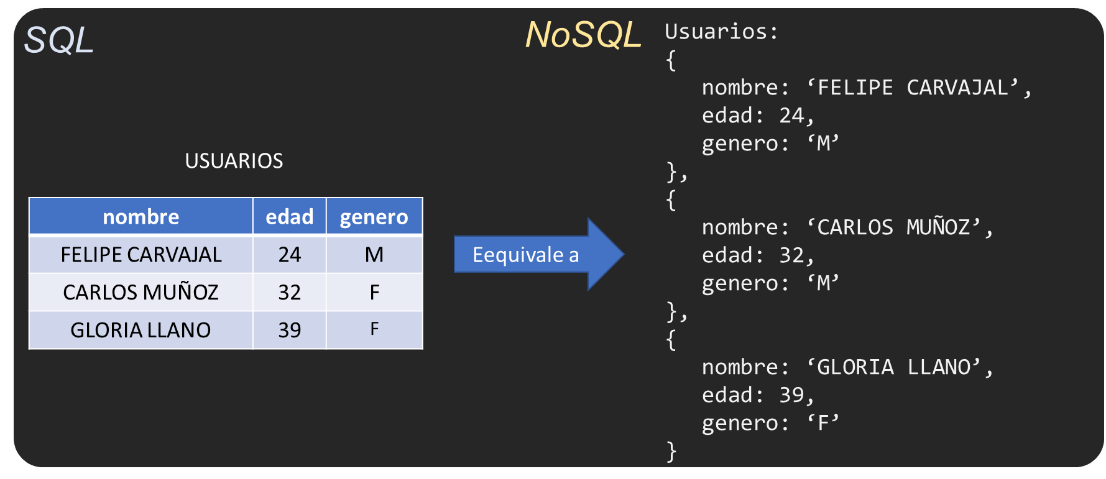
Es importante reconocer las diferencias entre las bases de datos relacionales SQL y las no relacionales (no SQL), para mejor entendimiento, por eso en la siguiente tabla se realiza un cuadro comparativo entre estos tipos de bases de bases de datos, empleando como ejemplo el motor de base de datos Mongo DB que actualmente es el gestor No SQL más usado y especializado en este tipo de conceptos de arreglo de datos.



Si bien las bases de datos NoSQL tienen una estructura conceptual distinta, podría verse en su equivalencia de la siguiente manera:

**Figura 9**

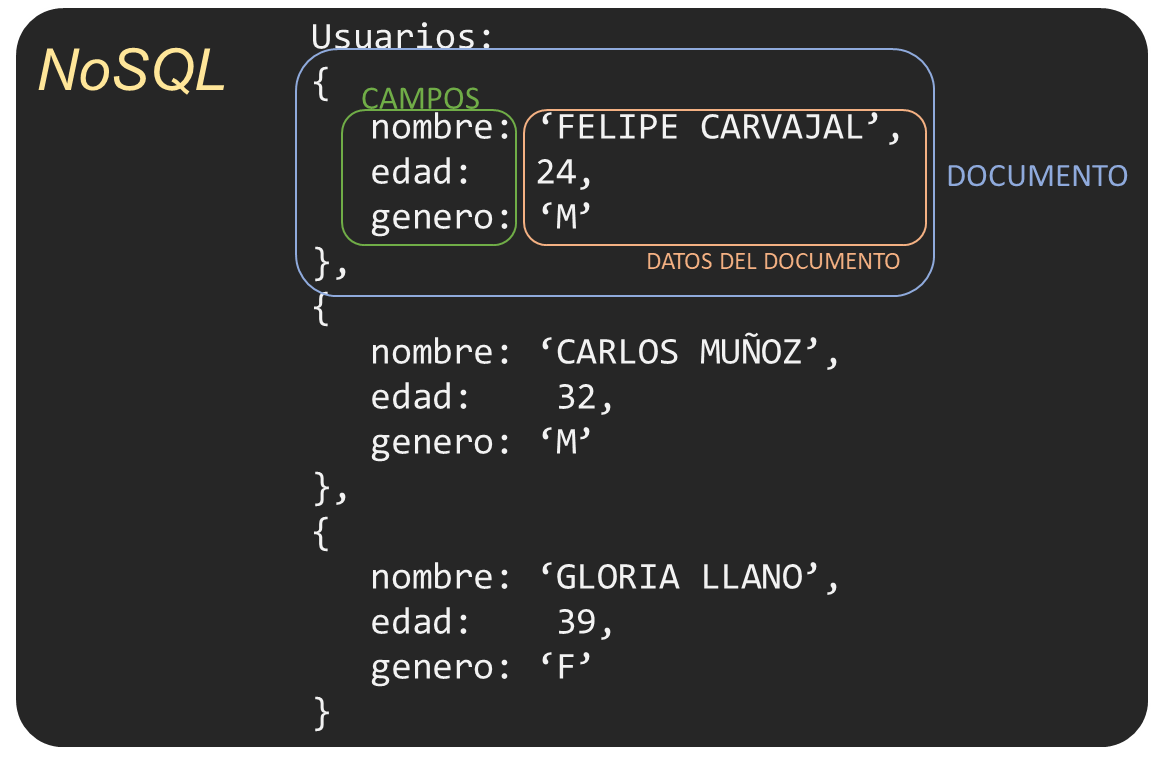
*Equivalencia de estructuras*



En síntesis, las columnas en NoSQL son campos, los registros de cada campo se llaman Datos del documento y las filas pasarían a denominarse Documento.

**Figura 10**

*Denominaciones de la estructura en NoSQL*

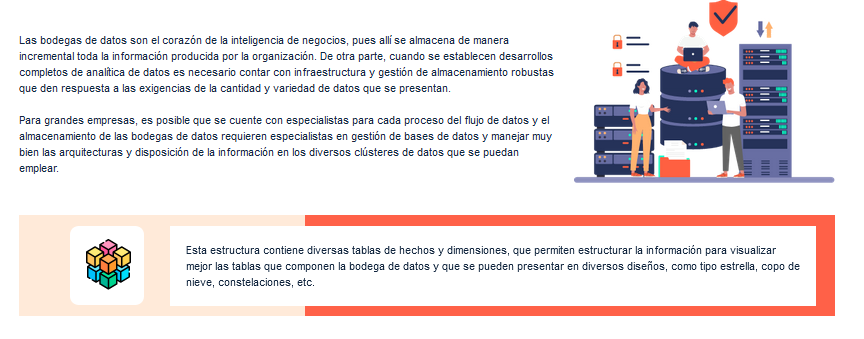


Al igual que otros lenguajes, para realizar consultas y gestiones y estructuras de datos es necesario instalar la herramienta necesaria, en este caso MongoDB, para realizar consultas y crear estructuras de datos. Aprender MongoDB es de gran utilidad, pues la tendencia del almacenamiento y gestión de grandes volúmenes de datos prefiere esta alternativa porque presenta grandes ventajas de desempeño para soluciones *big data*.

En el material complementario se relaciona material para profundizar en el aprendizaje de MongoDB.

## 4. Bodegas de datos





El fin de un proceso ETL será la de almacenar la información transformada en una o varias bases de datos destino; de esta manera, muchas pequeñas tareas y transformaciones se van almacenando de forma masiva y centralizada en una base de datos principal denominada *Data Warehouse* o bodega de datos.

### 4.1 Diseño de mercados

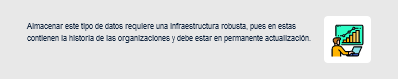
Uno de los aspectos más importantes a tener en cuenta son las tendencias tecnológicas y de consumo, pues a partir de ello los proyectos empiezan a tener aceptación y éxito. Se debe tener siempre presente qué está funcionando en la industria, cuáles son las prácticas de las grandes corporaciones, qué consume el público objetivo y cómo compra, para alinearse hacia estas tendencias buscando mejorar las probabilidades de éxito de las organizaciones.



### 4.2 Hechos, dimensiones

Las bodegas de datos están representadas en una base de datos que por lo general son de tipo SQL con un diseño Dimensional, es decir, cada *Datamart* cuenta con una tabla de hechos y otras tablas a modo de catálogos que se denominan dimensiones. A continuación.

* **Las tablas de hechos(*Fact*):** representan los eventos que suceden en determinado contexto-tiempo. Se caracterizan por permitir el análisis de los datos con el máximo detalle, son tablas que no tienen medida y suelen ser tablas más robustas, que contienen miles o millones de registros; además son las que más se actualizan. Por esta razón, cuando las transacciones en los sistemas OLTP son de manera masiva, se debe aplicar ingeniería de optimización de hechos, ya sea traer datos por periodo (*snapshot*), tablas agregadas, particionadas, etc.



* **Las tablas de dimensiones:** estas no suelen ser tan dinámicas como los hechos, en las dimensiones se recogen los puntos de análisis de un hecho. Por ejemplo, una venta se puede analizar respecto al día de venta, producto, cliente, vendedor, ciudad, entre otros, a su vez, estos elementos recién nombrados podrían categorizarse como dimensiones, como la dimensión tiempo, dimensión productos, dimensión cliente, etc.

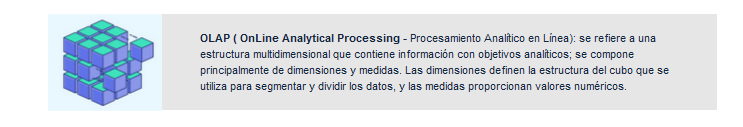
**Figura 11**

*Ejemplo estructura dimensional*



### 4.3 Cubos OLAP y ROLAP

Los cubos de datos no forman parte del proceso ETL, ya que estos son estructuras utilizadas por el Data Warehouse (DWH) para entregar datos que pueden ser consumidos a través de visualizaciones en tablas o gráficos multidimensionales.

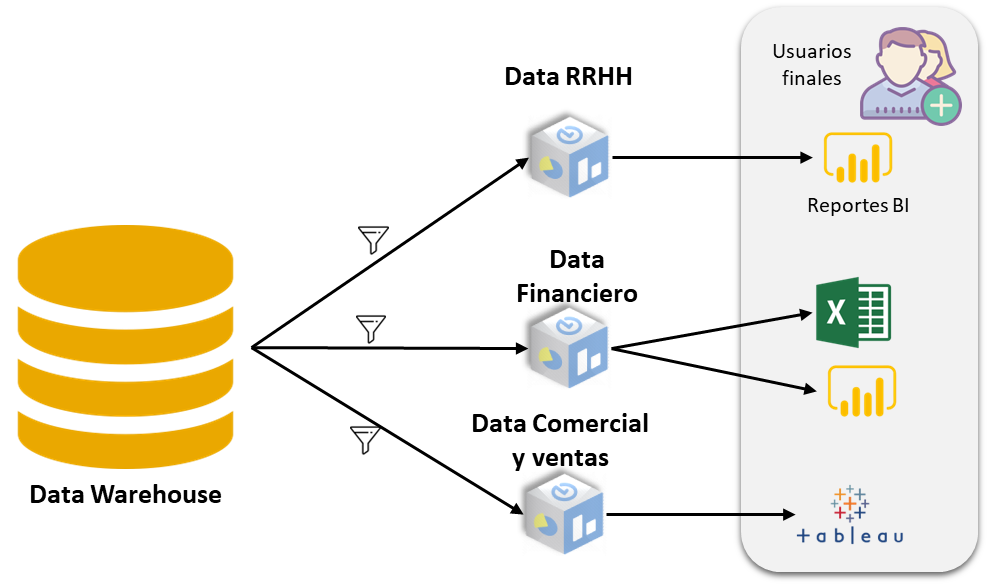


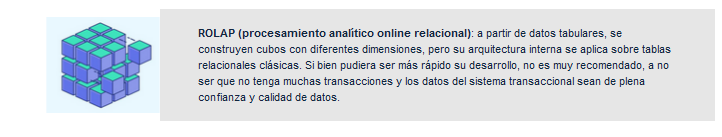
Como estructura lógica, un cubo permite a una aplicación cliente recuperar valores de medidas, como si estuvieran almacenados en las celdas del cubo, adicional se presentan estructuras por jerarquía de datos que podrían definir de alguna manera la profundidad o detalle de las consultas; el ejemplo más común de datos jerárquicos es el tiempo, que tiene año, trimestre, mes, semana, día; esto puede definir el detalle de los reportes.

Parte de la utilidad de los cubos de datos es que podrían ser consumidos por cada área del negocio, es decir, por cada departamento o área tener acceso a sus datos específicos y construir sus reportes en una aplicación local ya sea Excel, Power BI, Tableau u otro disponible.

**Figura 12**

*Autoconsumo de datos mediante OLAP*





### 4.4 Llenado de almacén de datos

Existen diversos conceptos del llenado de datos para la bodega de datos o almacén de datos (*Data Warehouse*), se presentan diversas arquitecturas, en otros materiales de formación se amplía los dos modelos predominantes: modelo Inmon y el modelo Kimball.

Independientemente del modelo empleado, todos los procesos ETL contienen los tres elementos principales: extracción,transformación y carga.

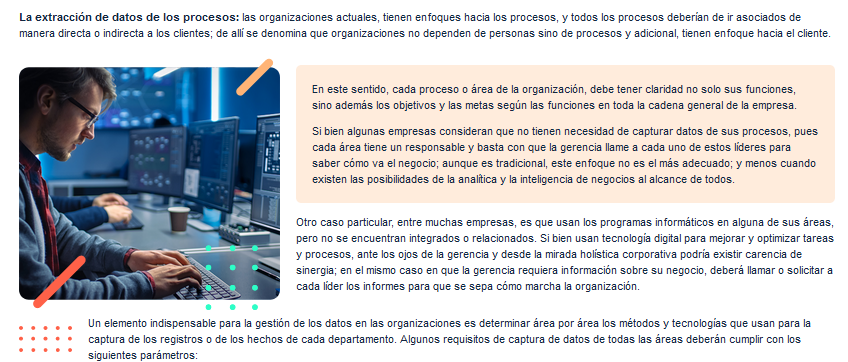
**Figura 13**

*Arquitectura inteligencia de negocios*

**

La arquitectura o estructura completa se puede resumir en dos grandes procesos: el proceso ETL y el desarrollo de reportes con los cuales los datos toman valor y conocimiento para el negocio.

**ETL:** consta de tres etapas, el desarrollador de inteligencia de negocio deberá tener en cuenta la extracción, identificando las fuentes de datos de los sistemas OLPT y otros no transaccionales, pero que aportan al conocimiento del negocio, así como su transformación y carga en bases de datos robustas que conservan las dimensiones, tablas de hechos y medidas de pertinentes.



Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

No basta con tener registrados estos cinco mínimos componentes de las tareas más relevantes para que los datos muestran una radiografía clara sobre cada proceso. Además, es importante tener claridad sobre la calidad de los datos para que se puedan integrar y tener plena confianza.

Escala de tiempo

Descripción generada automáticamente

En términos generales, la gerencia ya no tendrá que llamar a los líderes de cada proceso para que les dé explicación de una situación general, ya las directivas tendrán las cifras en sus teléfonos celulares, en cualquier lugar a cualquier hora. Esto permite tener empresas más eficientes, con capacidad de reaccionar más pronto y tomar mejores decisiones.

## 5. Herramientas tecnológicas de ETL

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza baja

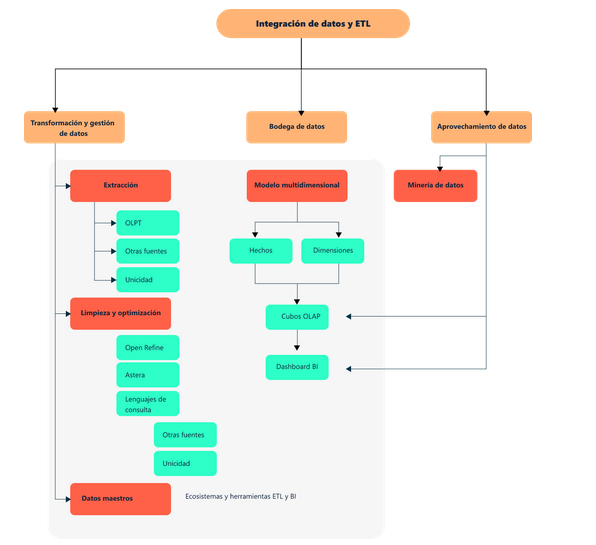
Las marcas más destacadas actualmente y que sirven para implementaciones completas de soluciones analíticas son:



En el mercado existen más de un centenar de posibilidades, es un sector de alto crecimiento y sus herramientas se van siendo más especializadas, integrándose con IA, grandes capacidades de infraestructura y facilidad de manejo.

1. **Síntesis**

A continuación, se presenta el diagrama que representa el resumen de las temáticas que están desarrolladas en el componente formativo, en donde se explica el proceso de integración de datos y ETL, que comienza con la extracción de datos de diversas fuentes, seguido de la limpieza y optimización con herramientas como Open Refine y Astera.



1. **ACTIVIDADES DIDÁCTICAS**

|  |  |
| --- | --- |
| DESCRIPCIÓN DE ACTIVIDAD DIDÁCTICA | |
| Nombre de la Actividad | Autoevaluación de conceptos sobre el proceso de integración de datos y ETL. |
| Objetivo de la actividad | Identificar los conceptos sobre proceso de integración de datos y ETL con el fin de reforzar los conocimientos adquiridos en el componente. |
| Tipo de actividad sugerida |  |
| Archivo de la actividad  (Anexo donde se describe la actividad propuesta) | Anexo\_CF14\_ActividadDidactica |

**MATERIAL COMPLEMENTARIO:**

|  |  |  |  |
| --- | --- | --- | --- |
| Tema | Referencia APA del Material | Tipo de material  (Video, capítulo de libro, artículo, otro) | Enlace del Recurso o  Archivo del documento o material |
| 1. Extracción y minería de datos | Conesa, C., J., y Curto, D., J. (2013). *Introducción al Business Intelligence.* Editorial UOC. | Video | <https://youtu.be/FJ91HT6aNiM> |
| 2.2 Astera | Astera software. (2020). *Extracción de PDF y exportación a Excel en Astera ReportMiner.* | Blog | <https://www.astera.com/es/tipo/blog/extraer-datos-de-pdf-a-excel/> este es en español |
| 3.2 SQL | **L**earn SQL: The best & easiest way to learn SQL. (s. f.-a). SQL Easy**.**. | Blog | <https://www.sql-easy.com/es/> |
| 3.3 NO-SQL | Canal Ecosistema de Recursos Educativos Digitales SENA. (2021). *NOSQL.* | Video clase | <https://youtu.be/u1IKJMISMgs> |

1. **GLOSARIO:**

|  |  |
| --- | --- |
| TÉRMINO | SIGNIFICADO |
| *Bytes* | unidad de medida de información. 1 *byte* corresponde a 8 bits, y a partir de esta unidad se determina el volumen de la información. |
| *Dashboard* | tableros de mando, es el recurso que resulta a partir del proceso de ETL. Es la manera de consumir datos y proporcionar conocimiento del negocio |
| *Datamart* | es la versión específica de cada área del *Data Warehouse*, son los datos concentrados por cada área del negocio. Son subconjuntos de colección de datos que alimentan a la bodega de datos y el resto de los recursos analíticos |
| IA | abreviación de inteligencia artificial. Área informática que simula procesos cognitivos humanos tales como aprendizaje, decisiones, y procesos complejos. |
| IDE | *Integrated Development Environment*: entorno de desarrollo integrado; se trata de una herramienta o entorno que integra otras herramientas, de esta manera el desarrollador no se preocupa de instalar recursos adicionales, todo estará en una sola herramienta que integra otras para que así, el profesional se dedique solo a la programación |
| *Machine Learning* | área de la IA que se responsabiliza de procesos de aprendizaje en el contexto de los datos se establecen aprendizaje supervisado y no supervisado, dependiendo del modelo de aprendizaje se establecen los algoritmos para desarrollar modelos predictivos y prescriptivos según el modelo analítico. |

1. **REFERENCIAS BIBLIOGRÁFICAS:**

Curto, D., J. (2016). *Introducción al business intelligence.* Editorial UOC. <https://elibro-net.bdigital.sena.edu.co/es/lc/senavirtual/titulos/101030>

Gorenés, R., J., Casas, R., J., y Minguillón, A., J. (2017). *Minería de datos: modelos y algoritmos.* Editorial UOC. <https://elibro-net.bdigital.sena.edu.co/es/lc/senavirtual/titulos/58656>

Pang, A., Markovski, M., & Ristik, M. (2022). *Los 10 principales proveedores de software de análisis y BI, tamaño del mercado y pronóstico del mercado 2021-2026.* Apps Run the World. <https://www.appsruntheworld.com/top-10-analytics-and-bi-software-vendors-and-market-forecast/>

Stibo System MDM. (2019). *¿Qué es la gestión de datos maestros?* Stibo System. <https://www.stibosystems.com/es/what-is-master-data-management>

1. **CONTROL DEL DOCUMENTO**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Nombre | Cargo | Dependencia | Fecha |
| Autor (es) | Jaime Hernán Tejada | Experto Temático | Regional Norte de Santander- Centro CIES | Noviembre de 2022 |
| Giovanna Andrea Escobar Ospina | Diseñador Instruccional | Regional Norte de Santander- Centro CIES | Noviembre de 2022 |
| Silvia Milena Sequeda Cárdenas | Asesora pedagógica y metodológica | Regional Distrito Capital - Centro de Diseño y Metrología | Noviembre de 2022 |
| Rafael Neftalí Lizcano Reyes | Responsable Desarrollo Curricular | Regional Santander - Centro Industrial del Diseño y la Manufactura. | Noviembre de 2022 |
|  | José Gabriel Ortiz Abella | Corrector de estilo | Regional Distrito Capital - Centro de Diseño y Metrología. | Diciembre del 2022. |

1. **CONTROL DE CAMBIOS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Nombre | Cargo | Dependencia | Fecha | Razón del Cambio |
| Autor (es) | Sandra Paola Morales Paez | Evaluador Instruccional | Regional  Santander  Centro Agroturístico | 3 de junio de 2024 | Adecuaciones a 2024 |
| Claudia Johanna Gómez Pérez | Responsable Línea Santander |