

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.

Tabla de contenido

Introducción	1
1. Extracción y minería de datos	2
2. Técnicas de limpieza y transformación de datos.....	5
2.1. Open Refine	7
2.2. Astera	8
3. Modelo de datos transaccionales	10
3.1. OLTP.....	12
3.2. SQL	16
3.3. No-SQL.....	22
4. Bodegas de datos.....	25
4.1. Diseño de Mercados	26
4.2. Hechos, dimensiones	27
4.3. Cubos OLAP y ROLAP	29
4.4. Llenado de almacén de datos	30
5. Herramientas Tecnológicas de <i>ETL</i>	35
Síntesis	39
Material complementario.....	40
Glosario	40

Referencias bibliográficas	41
Créditos	42

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:

Video 1. Proceso de integración de datos y *ETL*



[Enlace de reproducción del video](#)

Síntesis del video: Migración Proceso de integración de datos y *ETL*

El proceso de integración de datos *ETL* (Extracción, Transformación y Carga) es fundamental en las disciplinas informáticas, siendo central para la gestión y toma de decisiones en las organizaciones. Día tras día, se producen datos de manera dispersa e independiente, pero estos por sí solos no aportan valor agregado; por el contrario,

generan costos de almacenamiento sin el aprovechamiento que pudieran tener. Una vez se tengan claras las preguntas del negocio claves y estratégicas, se podrá definir el origen de los datos y el flujo para los procesos analíticos que permitirán procesar toda la información de manera estructurada, vinculando los recursos tecnológicos y transformando los datos en conocimiento. Cuando se realiza una adecuada identificación de las necesidades para el análisis de información, se podrá realizar seguimiento al rendimiento empresarial, las tendencias de mercado y las oportunidades comerciales. De esta manera, las empresas podrán tomar decisiones más inteligentes. En el componente formativo, se abordan los conceptos para la extracción y minería de datos, además de realizar análisis exploratorios de datos y el reconocimiento de las herramientas para *ETL*.

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 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 necesarias, no son del todo suficientes para llegar hasta el conocimiento que se esconde detrás de

estructuras y relaciones poco superficiales en los datos (Gorenés, Casas y Minguillón, 2017).

La minería de datos se encuentra estrechamente relacionada con la aplicación de inteligencia artificial (IA), en específico con procesos y técnicas algorítmicas de *Machine Learning (ML)* con sus múltiples posibilidades que son tema de profundización en otros componentes, adicional se emplea otros aspectos de ingeniería como árboles de decisiones, redes neuronales, entre otros.

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.

La minería de datos como disciplina estudia las herramientas y algoritmos que le permitan acceder y recolectar datos a las organizaciones para que esa información sea predictiva y se facilite su análisis para la toma de decisiones ante situaciones presentadas, mejorar los procesos de las entidades y sirve para la toma de decisiones.

La gestión con la minería de datos se debe realizar considerando los siguientes elementos:

- **Identificar la información.** Encontrando las categorías o conjuntos de los datos.
- **Clasificar la información.** De forma automática encontrando los datos solicitados o de interés.

- **Describir o brindar conceptos o información.** De forma compilada utilizando el método adecuado.
- **Detectar anomalías en los datos.** Filtrando la información recibida para evitar desviaciones o fallas.
- **Detectar las variables de la información.** Encontrando cambios que pueden originar los datos.
- **Mapear la información para verificar su autenticidad.** Revisando la procedencia de la información desde sus orígenes.

Esta gestión se debe realizar a través de:

- **Proceso de datos.**
Se debe revisar la información y eliminar lo que no sirve.
- **Elección de características.**
Se deben tomar las variables influyentes para solucionar el problema.
- **Uso de algoritmos.**
Se debe utilizar el modelo de conocimiento más acorde según las técnicas.
- **Análisis y evaluación.**
Se debe comprobar que las validaciones sean reales.

En el siguiente recurso se expondrán las técnicas que utiliza esta herramienta de recolección de datos:

- **Inferencia estadística.**
Se utiliza para inducir por muestras estadísticas el actuar de una comunidad, así como sus características.
- **Árbol de decisión.**

Se plasma por medio de una imagen las decisiones que se pueden tomar ante diferentes situaciones y, así mismo, se grafican sus resultados o consecuencias.

- **Redes neuronales.**

Imitan las neuronas del cerebro dando variables predictivas para la toma de decisiones.

- **Inducción de reglas.**

Genera métodos y lineamientos automáticos para identificar la información.

- **Aprendizaje basado en instancias.**

Procesa los datos hasta brindar respuestas, mediado por situaciones o ejemplos previamente delineados.

- **Algoritmos genéticos.**

Brindan soluciones a una situación por medio de una serie de algoritmos que siguen una función genética del problema.

- **Programación lógica inductiva.**

Por medio de la inteligencia artificial, utiliza lógica, hipótesis y el conocimiento previo para encontrar alternativas de solución.

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.

Las organizaciones en la preparación de datos deben tener en cuenta:

- Claridad de la información y seguridad en el acceso de un lugar seguro.
- Confiabilidad de los datos obteniendo información fidedigna, veraz, visible y auditable por si se debe modificar.
- Comprensible y con capacidad de repetición, asegurando su entendimiento para generar estrategias.

Ahora, para preparar los datos estructurados se siguen las siguientes fases:

1. Identificación

Identificar los datos que se necesitan, adquirirlos, compilarlos y generar un canal o línea de acceso permanente.

2. Revisión

Revisar la calidad de los datos compilados, según la relación con la actividad o situación que se quiere solucionar.

3. Limpieza

Limpieza de la información que consiste en corregir, eliminar duplicidad, suprimir datos incompletos y salvaguardar información confidencial.

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.

Como complemento, puede visitar el siguiente enlace:

- **Funcionalidades *Open Refine***

Configuración básica de la interfaz de usuario de *Open Refine*

[Ir al sitio](#)

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

ID	NOMBRE	APELLIDO	EDAD	GÉNERO		ID	NOMBRES_COMPLETOS	EDAD	GÉNERO
1	Felipe	Carvajal	24	M	Limpieza de datos →	1	FELIPE CARVAJAL	24	M
2	Carlos	Muños	32	F		2	CARLOS MUÑOZ	32	F
3	Gloria	llano	39	Femenino		3	GLORIA LLANO	39	F
4	Eugenio	Gómez	24	Hombre		4	EUGENIO GÓMEZ	24	M
5	Amparo	Martínez	54	Fem		5	AMPARO MARTÍNEZ	54	F
6	OSCAR	Pérez	33	OTRO		6	OSCAR PÉREZ	33	OTRO

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.

Enfocándose en el proceso de refinamiento, al cual se denomina *wrangle*, usa su herramienta llamada *ReportMiner*, la cual suele ser muy buena, entre otras características, permite incluso extraer y hacer *wrangle* a archivos PDF y otras fuentes de datos.

En apartados siguientes, cuando se hable de herramientas para *ETL*, se mencionará de nuevo esta herramienta que, sin duda, debe estar entre el ramillete de

opciones para las empresas que deseen entrar en la tendencia de la inteligencia de negocio.

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***).

Los datos maestros son una arquitectura para administrar, centralizar, organizar, clasificar, localizar, sincronizar y enriquecer los datos según las reglas de negocio. La gestión de datos maestros (MDM, por sus siglas en inglés) se resume en un repositorio central que garantiza una única visión autorizada de la información y optimiza costos e ineficiencias causadas por los almacenamientos de datos dispersos, apoya reportes de negocio mediante la ubicación exacta, vinculación y propiedades de entidades y de la información a través de productos, clientes, tiendas, ciudades, empleados, proveedores, activos digitales y más.

Por ello, se convierte en habilitador clave para proporcionar una vista única y confiable de la información empresarial crítica. Las fuentes de datos confiables ayudan a reducir los costos de integración de aplicaciones, mejoran la experiencia del cliente y generan información analítica accionable (*Stibo System MDM*, 2019).

Una solución de MDM procura superar algunos desafíos comunes en las organizaciones como:

- Silos de datos dispersos y múltiples versiones de sus datos.
- Datos errados como resultado de ingresos manuales y datos no validados.

- Costos en almacenamiento y seguridad en la información, tanto para su acceso, su conservación y disponibilidad.

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

Desde el contexto de inteligencia de negocios, los sistemas transaccionales son las fuentes de datos que alimentan los sistemas *ETL*, si bien existen otras fuentes tales como reportes de internet (análíticas de *web* y redes sociales, entre otros), archivos multimedia, archivos planos, tablas de Excel, etc.

Los sistemas transaccionales son las fuentes más convencionales en las fases de carga y transformación.

Los modelos de administración de datos se basan en ubicar donde se almacena la información, así como la extracción o consulta de la misma; estos son:

- **Por rango**

Se utiliza cuando se administra una gran cantidad de información y se hace como un árbol invertido donde la raíz es la fuente de datos y las hojas no producen información, tiene falencias cuando los datos son repetidos o redundantes.

- **Por red**

Se utiliza para resolver solución frente a gran volumen de datos evitando la redundancia, pero tiene fallas en el proceso de administración.

- **Multidimensional**

Se maneja para actividades directas y concretas de forma eficiente, brinda soluciones tipo relacional.

- **Por transacciones**

Se utiliza para remitir y recibir información de forma rápida y su uso debe contar con un buen sistema, ya que las operaciones que se generen deben ser al mismo tiempo, es decir, envío y recepción inmediata para garantizar el uso óptimo.

- **Por guía a objetos**

Este modelo hace uso de su administración y creación por funciones, lo que permite manejar programas y operaciones por separado, tiene la capacidad de trabajar sobre volúmenes altos de información de forma óptima.

- **Por relación**

Se utiliza para dar solución a situaciones reales y que la administración de los datos sea totalmente dinámica, tiene la facilidad de manejar volúmenes grandes de información de forma eficiente.

- **Por distribución**

Este modelo contempla su administración a través de la red en diferentes sitios, generando varias opciones para su edición y control; contiene alta capacidad de volumen de datos.

- **Por deducción**

Su administración se basa en situaciones presentadas, a través de la lógica que brinda las matemáticas, tiene la capacidad de almacenar y manejar volúmenes altos de datos.

3.1. **OLTP**

El procesamiento de transacciones en línea (**Online Transaction Processing**), en las gráficas típicas de los *ETL*, aparecen los *OLTP* como fuentes importantes de datos, esto se refiere a sistemas de información desarrollados o implementados en áreas específicas del negocio. Sirven como apoyo tecnológico específico de un área en particular; estas aplicaciones pueden ser de tipo administrativo en lo referente a las transacciones de las dependencias como Talento humano, área financiera, contable, nómina, etc.

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

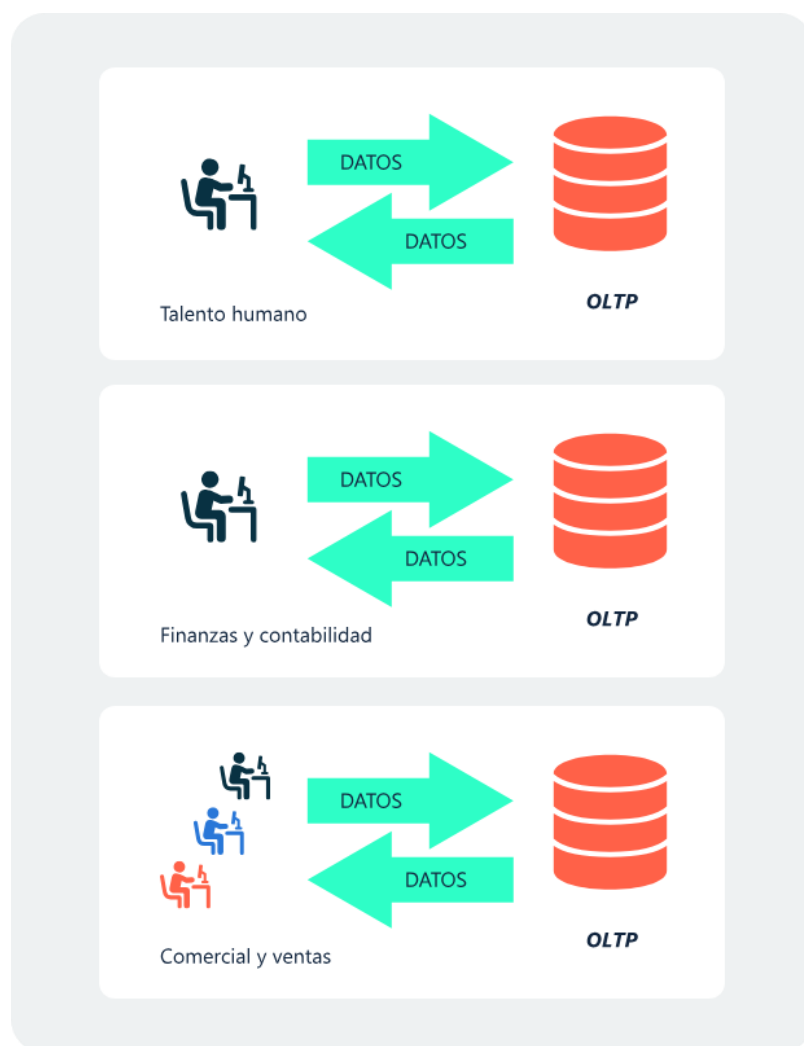
productos de infraestructuras y equipos eléctricos de gran envergadura, tendrá un área de mercadeo, otra de ofertas, licitaciones y ventas.

Si bien pueden ser áreas separadas, tienen entre sí una cadena que debe estar bien tejida y coordinada para mejorar los procesos; por ejemplo, el departamento de mercadeo para detectar clientes potenciales con alta probabilidad de compra, estos solicitan la cotización de un proyecto. El área de ofertas construye el proyecto y el camino ideal es que el departamento de ventas cierre el negocio. De manera resumida es muy fácil explicarlo, sin embargo, no lo es, pues se debe tener registros de cada paso; es decir, desde las campañas de mercadeo, hasta el momento que el cliente pide cotización y luego la respuesta de la oferta para finalizar la fecha de la confirmación y cierre del negocio; pero no solo tiene que ver con momentos y acciones, también se involucran personas para saber quién es el responsable, identificar plenamente el cliente y otros detalles.

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

Figura 3. Sistemas transaccionales *OLTP*



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:

Hojas de cálculo

Este tipo de información, aunque suele ser compleja en su calidad de datos y consistencia, no se deben descartar y, si es el caso, apoyar al proceso para optimizar su flujo de datos hacia el sistema BI. Para emplear estos recursos se hace necesario que exista un responsable que entregue o cargue de manera periódica los archivos actualizados para ir incrementando los datos.

Archivos planos

Por lo general, en CSV o TXT, suelen ser exportaciones de sistemas que no permiten conexión directa a las bases de datos.

Fuentes externas

Dependiendo de la solución, se hace necesario cruzar datos abiertos, ya sea para garantizar consistencias o para incluir datos de *big data*. Por ejemplo: traer información del clima de cada ciudad de operaciones podría incluirse en los reportes de analítica en caso de presentarse la hipótesis de que el clima influencia con las ventas.

Internet de las cosas (IoT)

Cada vez se usan sensores y dispositivos que registran acciones; los datos de estas tecnologías podrían incluirse en los sistemas de inteligencia de negocios para monitorear accesos, procesos de producción y otros controles que podrían estar también a disposición del nivel de decisión.

Analítica de internet

Los sitios *web*, redes sociales y comercios electrónicos tienen sus herramientas propias para llevar reportes sobre diferentes estadísticas y mediciones, estas

herramientas se pueden conectar a los tableros de mando o reportes que se realizan para la organización.

3.2. 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:

- ***DB Browser for SQLite.***

Aplicación liviana y muy fácil de instalar y manejar.

Sitio web. [Ir al sitio](#)

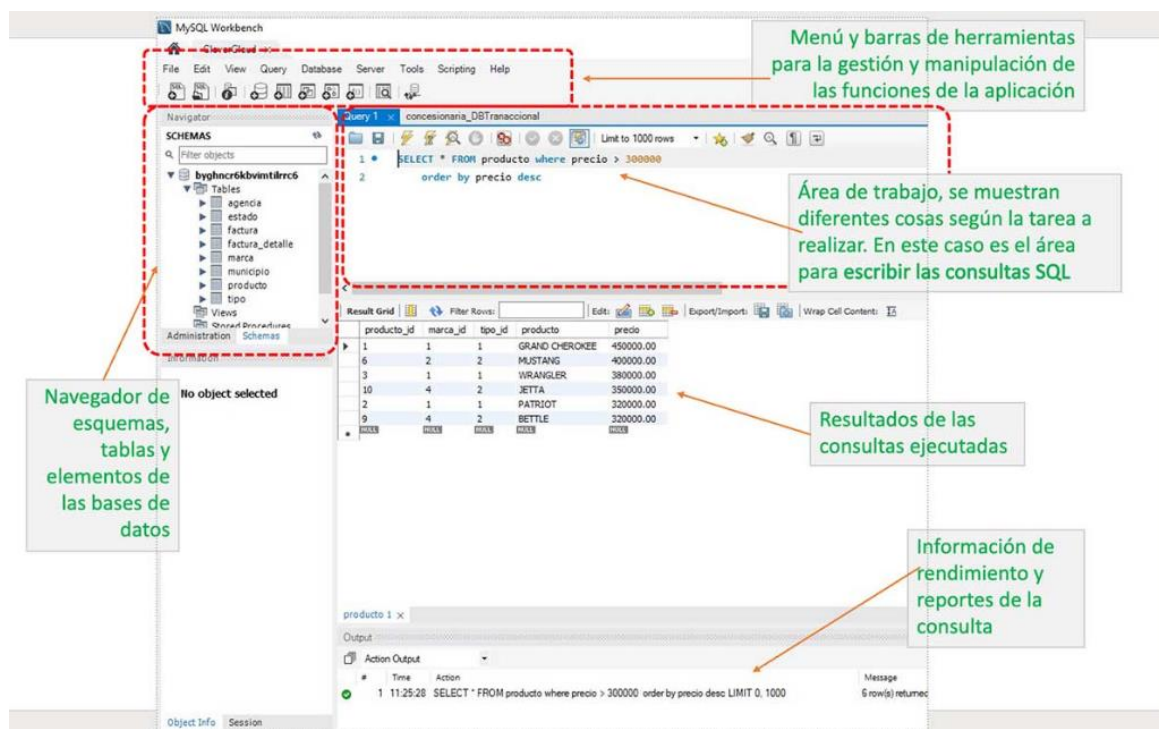
- ***MySQL Workbench***

Es una herramienta muy popular entre los desarrolladores, es de uso libre y presenta muy buenas funciones para la gestión.

Sitio web. [Ir al sitio](#)

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:

- **CREATE TABLE** crear una tabla.
- **NULL** para indicar valor nulo (ninguno).
- **UNIQUE** limita a un único valor en un campo.
- **PRIMARY KEY** asigna identificadora un campo.
- **DROP** eliminar tabla o base de datos completa.
- **TRUNCATE** borrar datos completos de una tabla.

Las sentencias básicas de consulta:

- **SELECT** para llamar datos.
- **WHERE** para incluir condiciones.

- **ORDER BY** ordenar los resultados.
- **INSERT INTO** para adicionar o insertar datos.
- **UPDATE** para actualizar datos.
- **DELETE** para borrar datos.

Las sentencias avanzadas.

- **LIMIT** el número de registros de consulta.
- **LIKE** busca en datos alfanuméricos por patrones dados.
- **JOIN** unir o combinar datos de diferentes tablas.

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.

- **Programiz Online SQL Editor**

Ingresar a la plataforma:

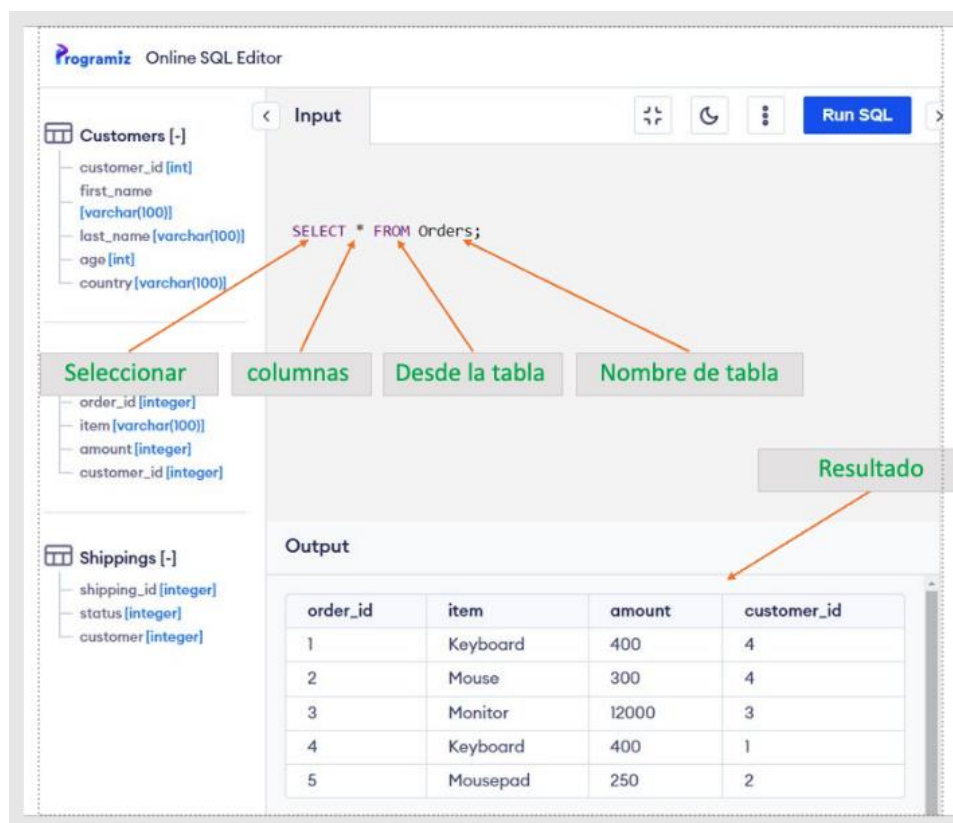
[Ir al enlace](#)

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

Sentencia: *Select*.

Para iniciar, el ejemplo sencillo de presentar todas las columnas de la tabla *Orders*.

Figura 5. Sentencia simple de SQL

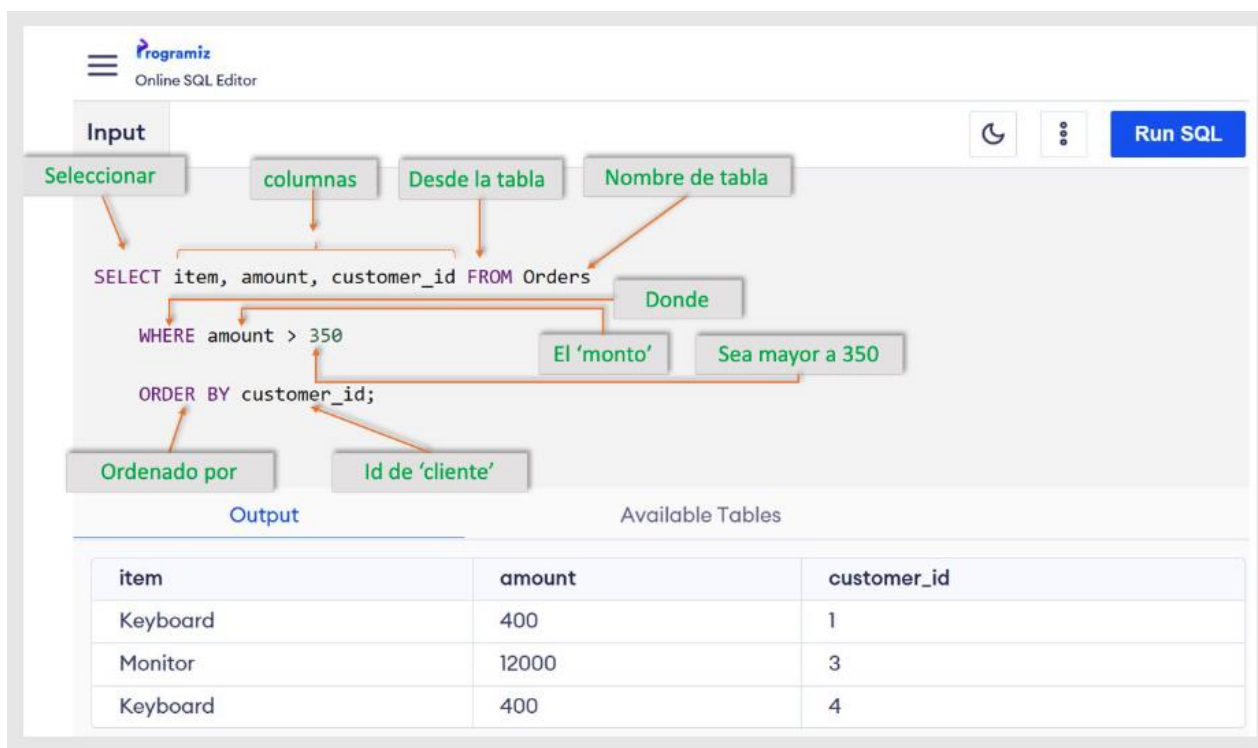


The screenshot shows the Programiz Online SQL Editor interface. On the left, there are two database schemas: 'Customers' and 'Shippings'. The 'Customers' schema includes columns: customer_id [int], first_name [varchar(100)], last_name [varchar(100)], age [int], and country [varchar(100)]. The 'Shippings' schema includes columns: shipping_id [integer], status [integer], and customer [integer]. The main 'Input' area contains the SQL query: `SELECT * FROM Orders;`. Four orange arrows point from labels below to parts of the query: 'Seleccionar' points to 'SELECT', 'columnas' points to '*', 'Desde la tabla' points to 'FROM', and 'Nombre de tabla' points to 'Orders'. A 'Run SQL' button is in the top right. Below the input, the 'Output' section displays a table with 5 rows and 4 columns. An orange arrow labeled 'Resultado' points to the output table.

order_id	item	amount	customer_id
1	Keyboard	400	4
2	Mouse	300	4
3	Monitor	12000	3
4	Keyboard	400	1
5	Mousepad	250	2

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



Programiz Online SQL Editor

Input

Seleccionar columnas Desde la tabla Nombre de tabla

SELECT item, amount, customer_id FROM Orders

Donde

WHERE amount > 350

El 'monto' Sea mayor a 350

ORDER BY customer_id;

Ordenado por Id de 'cliente'

Output

item	amount	customer_id
Keyboard	400	1
Monitor	12000	3
Keyboard	400	4

Available Tables

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

Figura 7. Insertar registro en SQL



Input

Run SQL

Available Tables

Customers

customer_id first_name last_name age country

1 John Doe 31 USA

2 Robert Luna 22 USA

3 David Robinson 22 UK

4 John Reinhardt 25 UK

5 Betty Doe 28 UAE

7 Oscar Gómez 27 COL

Registro agregado

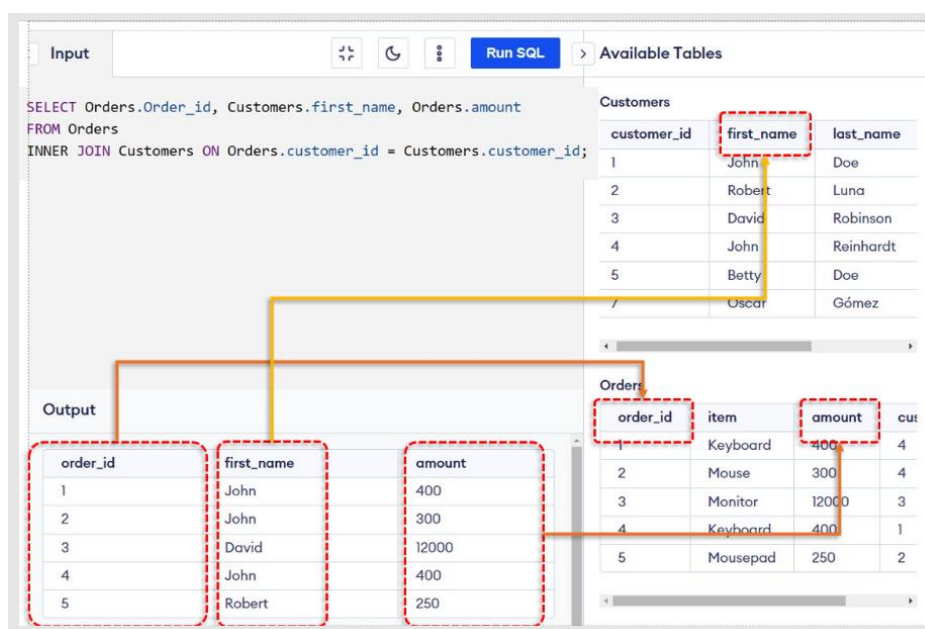
Para unir tablas o traer datos de otras tablas, es necesario tener claridad sobre las columnas que relacionan una tabla a otras, pues las uniones solo son posibles si se tiene un campo en común que asocie los conjuntos de datos.

En el siguiente ejemplo se tienen las tablas *Orders* (órdenes) y *Customers* (clientes), el campo en común en la tabla ordenes que asocia un cliente a través de “*customer_id*”.

A continuación, se desea ver en una sola tabla el id de la orden y el monto (que están en la tabla ‘*Orders*’), pero también el nombre del cliente (que está en la tabla *Customers*)

En el proceso *ETL* se presentan necesidades que en el ambiente gráfico se queda corto y será imposible de realizar, si bien la mayoría de las herramientas son basadas en gráficos y ventanas, se hace necesario codificar algunas tareas del proceso de carga y transformación, por esto es importante tener claridad y habilidades sobre el manejo especialmente de *SQL* y otros lenguajes de consulta de datos.

Figura 8. JOIN - Uniones en SQL



Input

```
SELECT Orders.Order_id, Customers.first_name, Orders.amount
FROM Orders
INNER JOIN Customers ON Orders.customer_id = Customers.customer_id;
```

Run SQL

Available Tables

Customers

customer_id	first_name	last_name
1	John	Doe
2	Robert	Luna
3	David	Robinson
4	John	Reinhardt
5	Betty	Doe
7	Oscar	Gómez

Orders

order_id	item	amount	cur
1	Keyboard	400	4
2	Mouse	300	4
3	Monitor	12000	3
4	Keyboard	400	1
5	Mousepad	250	2

Output

order_id	first_name	amount
1	John	400
2	John	300
3	David	12000
4	John	400
5	Robert	250

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

- **SQL Joins**

En este sitio indican cómo usar cada sentencia SQL, para que seguir profundizando en el tema: [Ir al enlace](#)

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 datos, empleando como ejemplo el motor de base de datos Mongo DB que actualmente es el gestor *NoSQL* más usado y especializado en este tipo de conceptos de arreglo de datos.

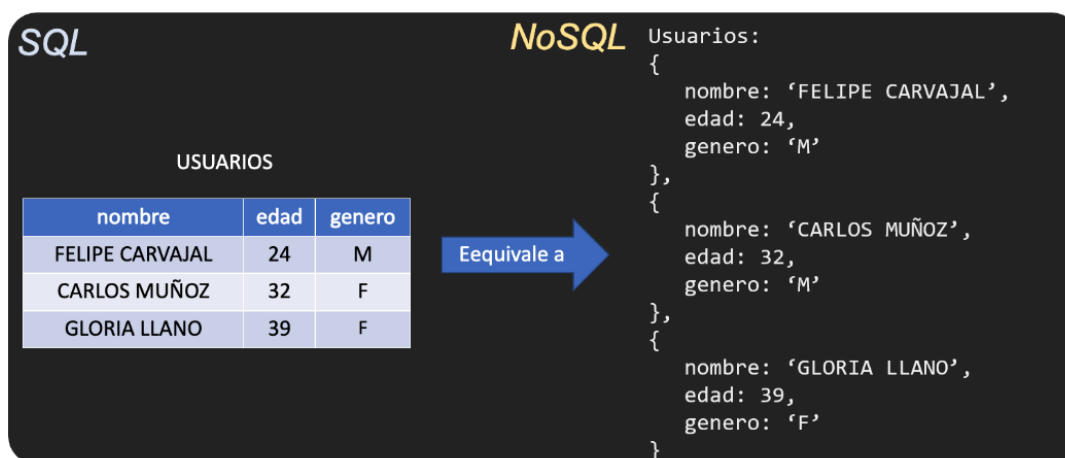
Tabla 1. Diferencias entre SQL y NoSQL

SQL	NoSQL
Relacionales: se asocia directamente otras colecciones de datos estructuradas.	No relacionales: no es requisito asociarse con otras colecciones de datos.
Estructura lenguaje SQL.	Estructura JavaScript.

SQL	NoSQL
Tablas: usa índices para ordenar y ubicar datos.	Colecciones de documentos (<i>JSON</i> , <i>BSON</i>).
Las bases de datos se componen de tablas asociadas.	Se establece la base de datos como un objeto llamado colección que contiene documentos y estos a su vez pueden tener más objetos dentro, y tener otras colecciones embebidas.
Su velocidad se afecta a medida que incrementa los datos.	Es muy veloz, incluso con altos volúmenes de datos.
Los datos deben estar configurados estrictamente (tipo, tamaño, formato, etc.).	Es más libre. No es estricto con los esquemas.

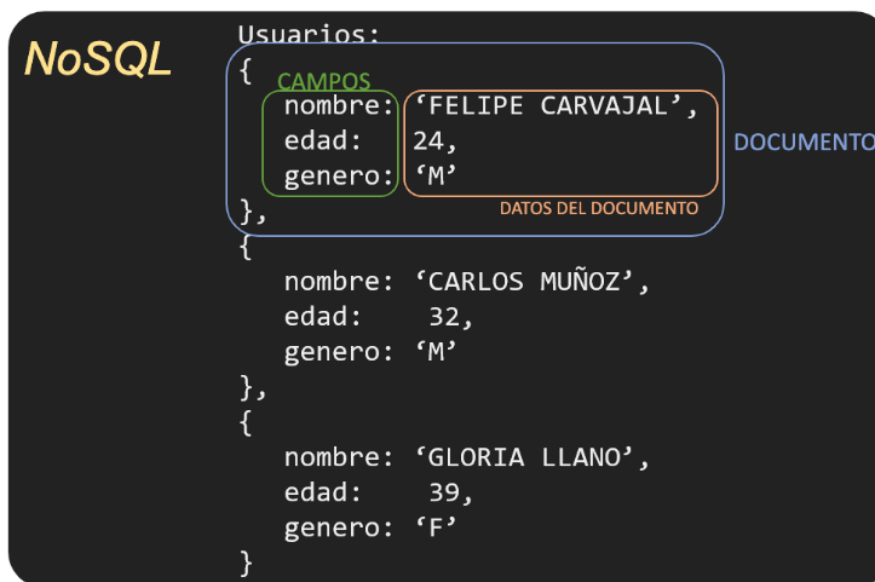
Si bien las bases de datos *NoSQL* tienen una estructura conceptual distinta, podría verse en su equivalencia de la siguiente manera:

Figura 9. Equivalencias 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

La información de las organizaciones proviene de fuentes de datos heterogéneas, muchas veces sin estar integradas, por lo cual tener reportes e información dispersa por áreas de la organización es un asunto que en la actualidad no debería presentarse, pues el fin de la inteligencia de negocios es facilitar el proceso y la disponibilidad de la información más relevante del negocio; para lograr esto se hace necesario centralizar los datos a través de bodegas de datos o *Data Warehouse (DWH)*.

La bodega de datos es una estructura diseñada y desarrollada para almacenar y procesar datos de múltiples fuentes y centralizadas para la elaboración de reportes y datos analíticos.

Las bodegas de datos son el corazón de la inteligencia de negocios, pues allí se almacena de manera incremental toda la información producida por la organización. De otra parte, cuando se establecen desarrollos completos de analítica de datos, es necesario contar con infraestructura y gestión de almacenamiento robustas que den respuesta a las exigencias de la cantidad y variedad de datos que se presentan.

Para grandes empresas, es posible que se cuente con especialistas para cada proceso del flujo de datos y el almacenamiento de las bodegas de datos requieren especialistas en gestión de bases de datos y manejar muy bien las arquitecturas y disposición de la información en los diversos clústeres de datos que se puedan emplear.

Esta estructura contiene diversas tablas de hechos y dimensiones, que permiten estructurar la información para visualizar mejor las tablas que componen la bodega de datos y que se pueden presentar en diversos diseños, como tipo estrella, copo de nieve, constelaciones, etc.

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.

Las tendencias y estar pendiente de las nuevas herramientas de los proveedores de computación en la nube ayudan a estar a la vanguardia en la implementación de herramientas tecnológicas y permanece competitivo alineándose a las necesidades de los nuevos clientes, cada día con más opciones por ofertas de la competencia haciéndolos al tiempo más y más exigentes pretendiendo soluciones ágiles y eficientes.

En el proceso mediante el cual la empresa u organización **obtiene, procesa y analiza la información de mercado o de la industria en la que compite** conformada por fuentes secundarias puede contener información clave de los clientes, los productos de

la competencia, la venta de estos productos, información sobre participación de mercados y perfiles de los clientes, o información específica sobre los posibles canales de distribución como tipos de establecimientos, ubicación geográfica, frecuencias de compra, tamaño o superficie, cantidades de compra con miras a usarla como soporte de sus planes de mercadeo y planeación comercial.

Esta información es elaborada por entidades públicas o privadas, instituciones académicas o empresariales que buscan ampliar el conocimiento del sector o brindar cifras estadísticas del comportamiento de las variables. Para el caso colombiano hay empresas como Nielsen, Meiko o Servinformación que se dedican a procesos de investigación de mercados para posteriormente comercializar esta información a empresas interesadas en adquirirla. Brindando información sobre georreferenciación, aspectos demográficos, psicográficos y conductuales.

4.2. Hechos, dimensiones

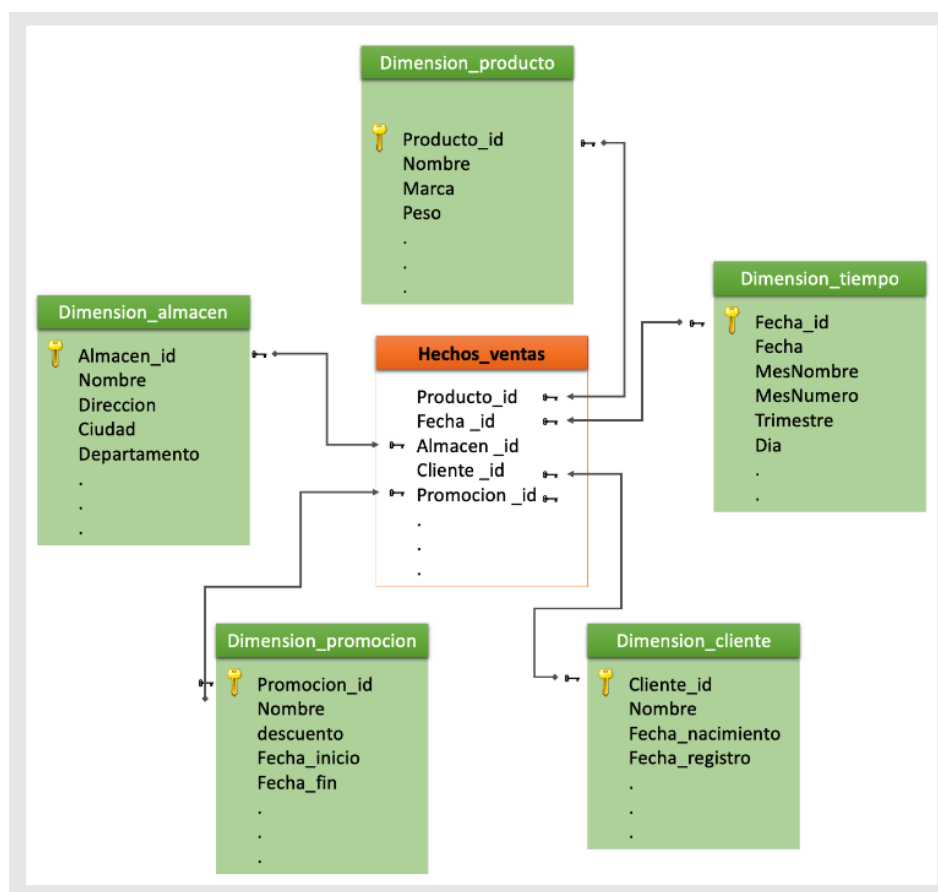
Las bodegas de datos están representadas en una base de datos que por lo general son de tipo *SQL* con un diseño Dimensional, es decir, cada *Datamart* cuenta con una tabla de hechos y otras tablas a modo de catálogos que se denominan dimensiones. A continuación.

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.

Almacenar este tipo de datos requiere una infraestructura robusta, pues en estas contienen la historia de las organizaciones y debe estar en permanente actualización.

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 son precisamente parte del proceso *ETL*, los cubos es una estructura en la que la bodega de datos o *el Data Warehouse (DWH)* entrega datos para consumirlos a través de visualización en tablas o gráficos multidimensionales.

OLAP (OnLine Analytical Processing - Procesamiento Analítico en Línea): se refiere a una estructura multidimensional que contiene información con objetivos analíticos; se compone principalmente de dimensiones y medidas. Las dimensiones definen la estructura del cubo que se utiliza para segmentar y dividir los datos, y las medidas proporcionan valores numéricos.

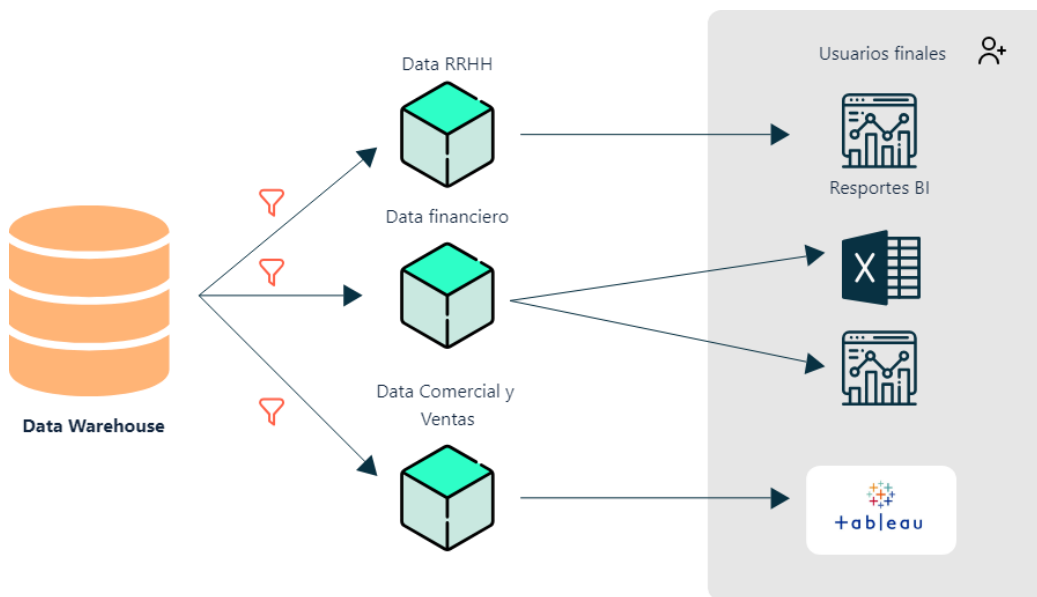
Como estructura lógica, un cubo permite a una aplicación cliente recuperar valores de medidas, como si estuvieran almacenados en las celdas del cubo, adicional se presentan estructuras por jerarquía de datos que podrían definir de alguna manera la profundidad o detalle de las consultas; el ejemplo más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

Las aplicaciones OLAP son uno de los pilares de cualquier solución de Inteligencia de Negocios, debido a que provee información sumariada a los que toman las decisiones, mediante métodos convenientes de navegación que les permiten analizar y mantener una conversación fluida con los datos de la organización, en óptimos tiempos de respuesta.

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*

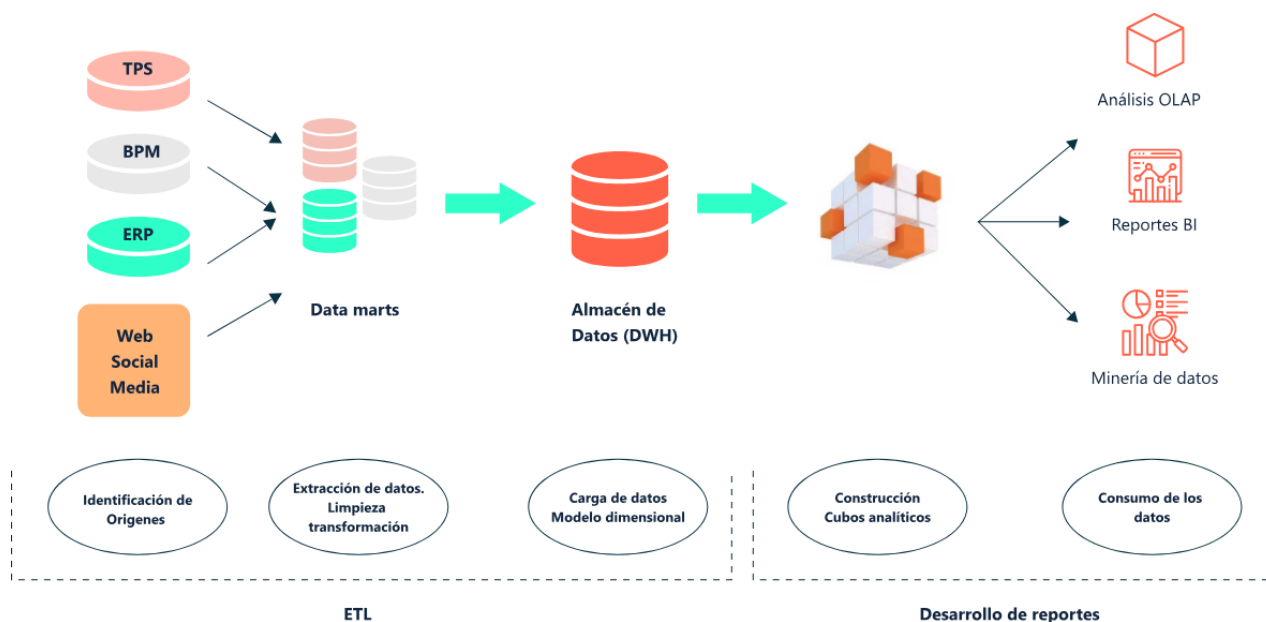


ROLAP (procesamiento analítico *online* relacional): a partir de datos tabulares, se construyen cubos con diferentes dimensiones, pero su arquitectura interna se aplica sobre tablas relacionales clásicas. Si bien pudiera ser más rápido su desarrollo, no es muy recomendado, a no ser que no tenga muchas transacciones y los datos del sistema transaccional sean de plena confianza y calidad de datos.

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.

La extracción de datos de los procesos: las organizaciones actuales, tienen enfoques hacia los procesos, y todos los procesos deberían de ir asociados de manera directa o indirecta a los clientes; de allí se denomina que organizaciones no dependen de personas sino de procesos y adicional, tienen enfoque hacia el cliente.

En este sentido, cada proceso o área de la organización, debe tener claridad no solo sus funciones, sino además los objetivos y las metas según las funciones en toda la cadena general de la empresa.

Si bien algunas empresas consideran que no tienen necesidad de capturar datos de sus procesos, pues cada área tiene un responsable y basta con que la gerencia llame a cada uno de estos líderes para saber cómo va el negocio; aunque es tradicional, este enfoque no es el más adecuado; y menos cuando existen las posibilidades de la analítica y la inteligencia de negocios al alcance de todos.

Otro caso particular, entre muchas empresas, es que usan los programas informáticos en alguna de sus áreas, pero no se encuentran integrados o relacionados. Si bien usan tecnología digital para mejorar y optimizar tareas y procesos, ante los ojos de la gerencia y desde la mirada holística corporativa podría existir carencia de sinergia; en el mismo caso en que la gerencia requiera información sobre su negocio, deberá llamar o solicitar a cada líder los informes para que se sepa cómo marcha la organización.

Un elemento indispensable para la gestión de los datos en las organizaciones es determinar área por área los métodos y tecnologías que usan para la captura de los registros o de los hechos de cada departamento. Algunos requisitos de captura de datos de todas las áreas deberán cumplir con los siguientes parámetros:

[Qué] Registrar hechos.

Todas las áreas deben registrar las tareas o acciones (compras, terminación de tareas, visitas, publicación de campañas, etc.).

[Cuándo] Registrar momentos.

Una de las características de la analítica es llevar una historia, no es posible que exista una historia si no hay fechas. Es importante determinar las fechas de cada hecho.

[Quién] Registrar responsable.

Cada tarea deberá tener un responsable, más cuando hay tareas entre diferentes áreas, de esta manera es posible identificar quién o en dónde se encuentran los embudos en los diferentes o procesos del negocio.

[Cuánto] Registrar cantidades.

La mayoría de los hechos, contienen un número que acompaña las características, ejemplo, una venta tiene detalles como cantidad de elementos y un monto de dinero, estas cifras siempre deberán registrarse; en otras tareas, como por ejemplo aprobar una solicitud, donde no hay un monto numérico si contiene un estado, por lo que también debe registrar los estados de estas tareas.

[A quién o de quién] Registrar clientes o proveedores.

Cada hecho o acción debe tener un cliente, si bien puede ser externo como las ventas también podría ser interno, por ejemplo, la aprobación de presupuesto, el cliente interno sería la dirección responsable de lo financiero; todos los hechos deben tener a quién para quién se le hace la acción.

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.

La transformación e integración de datos:

Cuando se tienen identificadas las fuentes de datos de cada área, se procede a extraer estos datos, se debe emplear metodologías para que los datos tengan una copia óptima en sistemas de almacenamiento centralizados y homogeneizados que garanticen la calidad de los datos y estén disponibles para la realización de reportes de todas las áreas.

Diseño de reportes:

Si bien estrictamente esta etapa no corresponde al proceso *ETL*, el objetivo es que todos esos datos recolectados, copiados y optimizados sean consumidos por el nivel de decisión, no necesariamente son los gerentes, los empleados también deben conocer sus gestiones y resultados en gráficos y *dashboards* que den cuenta de su propio rendimiento, allí también se genera una acción de autoevaluación y toman decisiones sobre sus propias funciones en la empresa.

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

Para la construcción de soluciones analíticas es fundamental seleccionar las herramientas adecuadas que permitan soportar cada uno de los procesos que se ejecutan en el ecosistema de procesamiento analítico. La oferta en el mercado de herramientas de inteligencia de negocios es muy alta, con algunos proveedores que representan corporaciones, tendencias tecnológicas y de negocio.

A la hora de decidirse por cuál ecosistema digital decidirse, es importante tener cuenta varios aspectos: entre otros, debe proporcionar herramientas *Front end* para usuarios con diferentes perfiles, debe permitir conexiones a diversas fuentes de datos, debe tener una infraestructura robusta y muy eficiente para el almacenamiento de *Data Warehouse* y *DataMart* que integre, además, componentes *ETL* como calidad de datos, diccionarios, repositorios centralizados y otras características importantes a la hora de elegir una herramienta o colección de herramientas.

No está de más recalcar que, más que las herramientas para el desarrollo de inteligencia de negocios, debe estar tras una estrategia clara y objetivos concretos para que el planteamiento de la arquitectura de soluciones analíticas tenga mayor éxito.

Herramientas para el desarrollo de BI

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

Tabla 2. Herramientas disponibles para soluciones analíticas

Marca	Sitio Web	Características
<i>SAS Institute</i>	https://www.sas.com/es_co/software/visual-analytics.html	Solución integral. Reportes interactivos. Descubrimiento visual. Analítica de autoservicio. Escalabilidad y gobierno desde un mismo entorno.
<i>Salesforce</i>	https://www.salesforce.com/mx/products/analytics/overview/	Tableau ha seguido creciendo como parte de <i>Salesforce</i> .
<i>SAP</i>	https://www.sap.com/latinamerica/products/technology-platform/bpc.html	Con muchos <i>softwares</i> corporativos, incorpora ahora herramientas de <i>ETL</i> y de inteligencia de negocios.

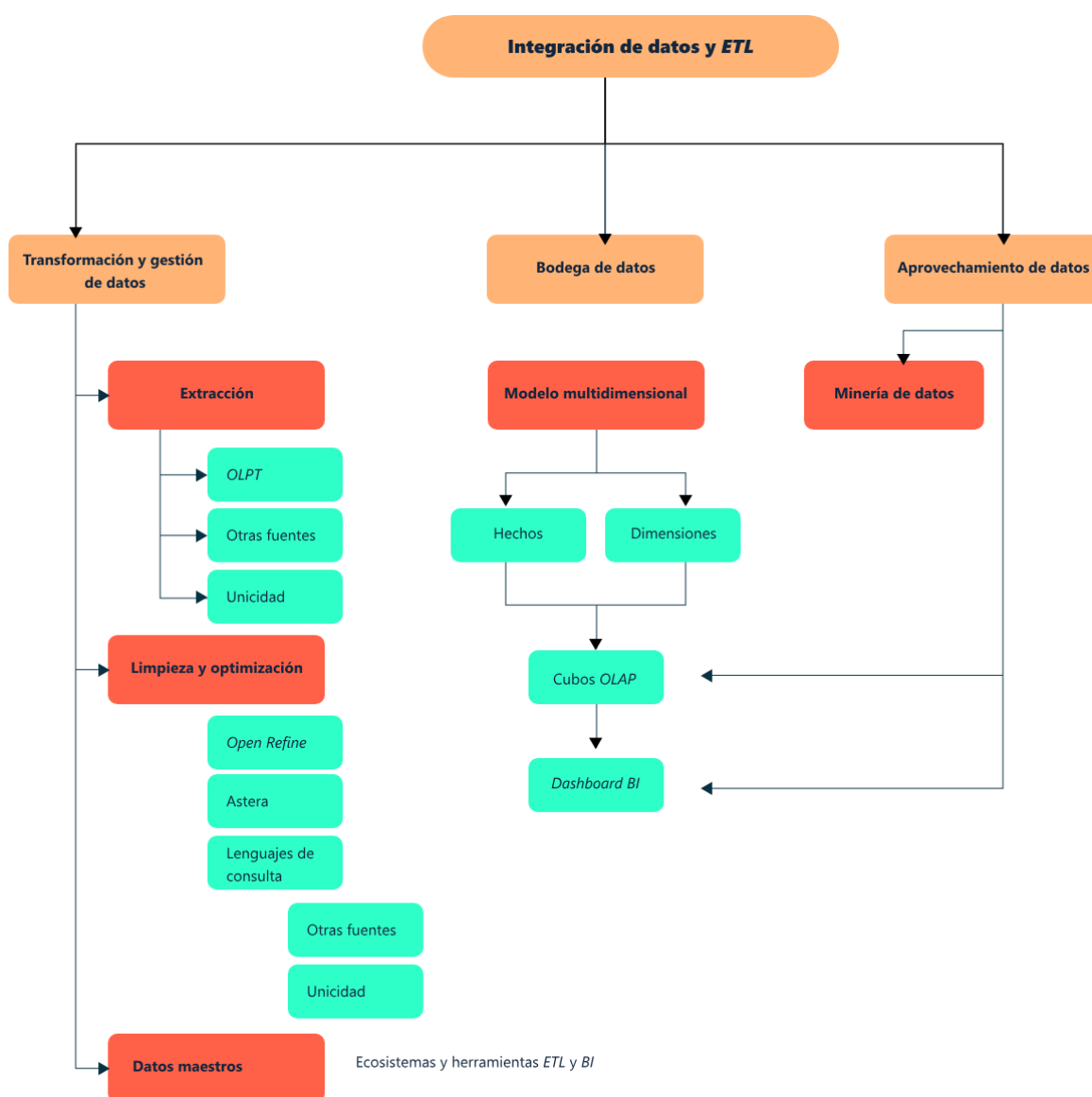
Marca	Sitio Web	Características
Microsoft	https://powerbi.microsoft.com/es-es/what-is-business-intelligence/	Visualizaciones de <i>Power BI</i> dentro de <i>Power Apps</i> y <i>Dynamics 365</i> aumenta las capacidades analíticas integradas de <i>Microsoft Power Platform</i> al permitir a los usuarios incorporar activos de inteligencia empresarial.
Oracle	https://www.oracle.com/lat/business-analytics/	<i>Oracle Analytics Cloud</i> se ha convertido en la herramienta analítica y de informes para las pequeñas empresas, más allá de las grandes corporaciones que lo usan.
Informática	https://www.informatica.com/products/cloud-data-integration.html	Integración de computación en la nube con herramientas de integración y procesos de datos hacia la inteligencia de negocios.
Quantexa	https://www.quantexa.com/platform/scoring-analytics/	Contextos completos de datos, incluye módulos de soluciones de inteligencia corporativa.

Marca	Sitio Web	Características
IBM	https://www.ibm.com/products/environmental-intelligence-suite	Anunciaron una solución comercial de gestión de datos empresariales en la nube híbrida con tecnología de datos de fuente abierta: <i>IBM Environmental Intelligence Suite</i> .
<i>Qlik</i>	https://www.qlik.com/es-es/products	Una gran marca que integra herramientas de infraestructura, bases de datos y analítica. Promete cerrar la brecha entre los datos, los conocimientos y la acción con <i>Qlik Cloud</i> ®, la única plataforma que reúne todos los datos y con su analítica.

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.

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.



Material complementario

Tema	Referencia APA del material	Tipo	Enlace
Extracción y minería de datos	Conesa, C., J., y Curto, D., J. (2013). Introducción al Business Intelligence. Editorial UOC.	Video	https://www.youtube.com/watch?v=FJ91HT6aNiM
Astera	Astera software. (2020). Extracción de PDF y exportación a Excel en Astera ReportMiner.	Blog	https://www.astera.com/es/tip-o/blog/extraer-datos-de-pdf-a-excel/
SQL	Learn SQL: The best & easiest way to learn SQL. (s. f.-a). SQL Easy.	Blog	https://www.sql-easy.com/es/
NO-SQL	Canal Ecosistema de Recursos Educativos Digitales SENA. (2021). NOSQL.	Video clase	https://www.youtube.com/watch?v=u1IKJMISMgs

Glosario

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.

D

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.

I

IA: abreviación de Inteligencia artificial. Área informática que simula procesos cognitivos humanos tales como aprendizaje, decisiones, y procesos complejos.

IDE: "*Integrated Development Environment*", en su traducción: 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.

M

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.

Referencias bibliográficas

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

Gorenés Roig, J., Casas Roma, J., & Minguillón Alfonso, J. (2017). Minería de datos: modelos y algoritmos. Barcelona: Editorial UOC. eLibro. <https://elibro-net.bdigital.sena.edu.co/es/lc/senavirtual/titulos/58656>

Pang, A., Markovski, M., & Ristik, M. (22 de septiembre de 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. (octubre de 2019). ¿Qué es la gestión de datos maestros?. Stibo system. <https://www.stibosystems.com/es/what-is-master-data-management>

Créditos

ECOSISTEMA DE RECURSOS EDUCATIVOS DIGITALES

Milady Tatiana Villamil Castellanos	Responsable del Ecosistema	Dirección General
Claudia Johanna Gómez Pérez	Responsable de Línea de Producción	Regional Santander - Centro Agroturístico

CONTENIDO INSTRUCCIONAL

Jaime Hernán Tejada	Experto Temático	Regional Norte de Santander- Centro CIES
Giovanna Andrea Escobar Ospina	Diseñador Instruccional	Regional Norte de Santander- Centro CIES
Silvia Milena Sequeda Cárdenas	Asesora pedagógica y metodológica	Regional Distrito Capital - Centro de Diseño y Metrología
José Gabriel Ortiz Abella	Corrector de Estilo	Regional Distrito Capital - Centro de Diseño y Metrología.
Rafael Neftalí Lizcano Reyes	Responsable Equipo de Desarrollo Curricular	Regional Santander – Centro Industrial del Diseño y la Manufactura
Sandra Paola Morales Páez	Evaluadora Instruccional	Regional Santander - Centro Agroturístico

DISEÑO Y DESARROLLO DE RECURSOS EDUCATIVOS DIGITALES

Julian Fernando Vanegas Vega	Diseñador de Contenidos Digitales	Regional Santander - Centro Agroturístico
Pedro Alonso Bolivar González	Desarrollador <i>Fullstack</i>	Regional Santander - Centro Agroturístico
Maria Alejandra Vera Briceño	Animadora y Productora Multimedia	Regional Santander - Centro Agroturístico

Lucenith Pinilla Moreno	Actividad Didáctica	Regional Santander - Centro Agroturístico
-------------------------	---------------------	---

VALIDACIÓN RECURSO EDUCATIVO DIGITAL

Laura Paola Gelvez Manosalva	Validadora de Recursos Educativos Digitales	Regional Santander - Centro Agroturístico
Erika Fernanda Mejía Pinzón	Validadora para Contenidos Inclusivos y Accesibles	Regional Santander - Centro Agroturístico