

# Modelamiento, análisis y preparación de datos

## Breve descripción:

El recurso educativo presenta contenidos y técnicas sobre conceptos, teorías y herramientas empleadas en sistematización de datos masivos. Se dan las pautas y contextos sobre los paradigmas más usados para la gestión de información enfocado a la analítica y carga masiva.

## Tabla de contenido

|   |    |
|---|----|
| Introducción .....  | 1  |
| 1. Tecnologías de información .....   | 2  |
| 1.1. Metodologías de diseño y normalización Principios de <i>ACID</i> ..... | 3  |
| 1.2. Técnicas de almacenamiento de datos y consultas .....                  | 5  |
| 1.3. Estándares técnicos .....  | 11 |
| 1.4. Ordenamiento de datos, indexación y recuperación .....                 | 12 |
| 2. Preparación de datos .....   | 18 |
| 2.1. Entendimiento de la <i>data</i> .....                                  | 21 |
| 2.2. Detección de errores y datos faltantes.....                            | 26 |
| 2.3. Identificación de variables importantes .....                          | 29 |
| 2.4. <i>Dataset</i> .....   | 31 |
| 3. La inteligencia de negocios .....  | 33 |
| 3.1. Identificación de las preguntas básicas .....                          | 34 |
| 3.2. Metodología de integración.....  | 35 |
| 3.3. Herramientas de administración.....                                    | 37 |
| 3.4. Técnicas de solución de problemas (modelación de datos).....           | 39 |
| 3.5. Metodologías de análisis ( <i>Kimball, Inmon</i> ).....                | 40 |
| 3.6. Verificación de valores y escalas .....                                | 42 |

|       |  |    |
|-------|--|----|
| 3.7.  | Procedimientos almacenados y funciones .....   | 45 |
| 3.8.  | Disparadores.....  | 45 |
| 4.    | Análisis exploratorio de datos .....   | 46 |
| 4.1.  | Estadística descriptiva y estadística inferencial .....  | 47 |
| 4.2.  | Población y muestra .....  | 47 |
| 4.3.  | Escalas de medida y clasificación de variable.....   | 48 |
| 4.4.  | Técnicas de análisis estadístico .....   | 50 |
| 5.    | Métodos para hacer análisis exploratorio de datos.....   | 52 |
| 5.1.  | Datos univariantes .....   | 53 |
| 5.2.  | Datos bivariantes .....  | 53 |
| 5.3.  | Datos multivariantes .....   | 53 |
| 5.4.  | Reglas de negocio .....  | 53 |
| 5.5.  | Tipo de restricciones.....   | 55 |
| 5.6.  | Programación transaccional.....  | 56 |
| 5.7.  | Programación de estructuras no lineales, desnormalización, series y<br><i>dataframes</i> ..... | 56 |
| 5.8.  | Álgebra relacional .....   | 57 |
| 5.9.  | <i>SQL</i> .....   | 61 |
| 5.10. | No <i>SQL</i> .....  | 66 |

|  |    |
|--|----|
| 5.11. <i>JSON, BSON y XML</i> .....                      | 68 |
| 5.12. <i>DDL, DML y DC</i> .....                         | 69 |
| 6. Estructuras y componentes de analítica de datos ..... | 70 |
| 6.1. Bodega de datos .....                               | 73 |
| 6.2. Tipos estrella .....                                | 74 |
| 6.3. Copo de nieve .....                                 | 75 |
| 6.4. Constelación .....                                  | 76 |
| 7. Herramientas para el análisis de datos.....           | 76 |
| 7.1. Entornos de desarrollo – <i>IDE</i> .....           | 78 |
| 7.2. Python .....  | 80 |
| 7.3. Librerías .....                                     | 81 |
| Síntesis .....   | 82 |
| Material complementario.....                             | 84 |
| Glosario .....   | 85 |
| Referencias bibliográficas .....                         | 88 |
| Créditos .....   | 90 |

## Introducción

En este componente se abordarán los conceptos y fundamentos de la inteligencia de negocios, con el fin de realizar el modelado, análisis y preparación de datos. Para una comprensión más detallada, se expone el siguiente video:

### Video 1. Modelamiento, análisis y preparación de datos



[Enlace de reproducción del video](#)

#### Síntesis del video: Modelamiento, análisis y preparación de datos

El modelamiento, análisis y preparación de datos son aspectos cruciales en la inteligencia de negocios, que impulsa el desarrollo empresarial a través del análisis, minería, predicción y visualización de datos. Estos datos, fundamentales en el ámbito empresarial y en la industria 4.0, se manejan con herramientas que muestran información histórica y actual en un contexto de negocios. La gestión adecuada de la

información es esencial para la transformación digital, utilizando minería de datos, *machine learning (ML)*, inteligencia artificial, ciencia de datos, tableros de mando y trabajo colaborativo, entre otros. Identificar adecuadamente las necesidades de análisis de información permite seguir el rendimiento empresarial, las tendencias de mercado y las oportunidades comerciales, lo que facilita decisiones empresariales más inteligentes. En el componente formativo, se abordan conceptos de inteligencia de negocios, modelamiento y preparación de datos, así como análisis exploratorios mediante métodos especializados, estructuración de componentes analíticos y reconocimiento de herramientas para el análisis de datos.

## **1. Tecnologías de información**

Las bases de datos son esa colección de datos integrados bajo un contexto o dominio, que contienen datos estructurados que reflejan relaciones, restricciones, validaciones y semánticas que reflejan las condiciones de negocio. Esta colección usa la computación para su almacenamiento y procesamiento.

Los sistemas de base de datos están conformados por los siguientes elementos básicos:

### **1. Hardware**

Consiste en los dispositivos y componentes electrónicos y mecánicos en los que se almacena, viaja y se conservan los datos (unidades de almacenamiento, red, procesadores, *RAM*, servidores, etc.).

## **2. Software**

Conjunto de programas, rutinas y comandos que permiten ejecutar tareas, tales como procesamiento de datos, lectura de datos y junto con el sistema operativo, se podrá administrar el *hardware*. Para las bases de datos existen los gestores de datos, ya sea SQL o NoSQL y las aplicaciones que usan los datos.

## **3. Los datos**

Son el objetivo principal, al ser información, se podría indicar que hace parte también del *software*. Su guardado lógico podría ser relacionales (SQL) y NO relacionales (NoSQL). De esto depende la manera en cómo funciona, cómo se procesa y su organización.

## **4. Personal**

Son los usuarios que están involucrados en el manejo y aprovechamiento de los datos, si bien hay varios niveles de usuarios, se podría dividir en dos principalmente en: no informáticos (que requieren la información para ejecutar tareas) y los Informáticos (los responsables del diseño y mantenimiento del sistema de bases de datos).

### **1.1. Metodologías de diseño y normalización Principios de ACID**

Antes de existir las bases de datos como se conocen hoy, las primeras gestiones en datos se basaban en procesar archivos, es decir, los registros se traducen en archivos que eran procesados por lotes gestionados por un sistema operativo. Incluso los datos andaban mezclados con los archivos de las aplicaciones, lo que suponía un gran inconveniente pues existía dependencia funcional de los archivos de las aplicaciones,

con los archivos del sistema operativo y los datos a guardar. Ello significaba que, si un dato se dañaba, era posible también que se estropeará la aplicación y viceversa.

**A mediados de la década de 1970**, se propone arquitecturas de diversos niveles para separar las capas según su funcionalidad, así que los Sistemas de Gestión de Bases de Datos (*DBMS Data Base Management System*) tienen sus capas según el contexto y su funcionamiento es independiente de los sistemas de gestión, de relación de datos, de la conservación de los datos, la interfaz gráfica, etc.

Todas las bases de datos deben tener unos atributos mínimos que garanticen que las transacciones se ejecuten de manera confiable. Entiéndase transacción como una unidad compuesta por varias tareas, cuyo resultado final debe exigir que se ejecuten todas o ninguna de ellas (Pulido Romero, Escobar Domínguez, & Núñez Pérez, 2019).

Un ejemplo frecuente es una App bancaria, la cual se usa para realizar compras, esta acción encarna una cantidad de tareas internas, de manera básica deberá restar de mi cuenta el valor del producto a pagar y a su vez se deberá sumar a la cuenta del vendedor; sería un gran problema si en la acción se debita de la cuenta del comprador y existe una falla del sistema que evita se sume al saldo del vendedor. Para que las transacciones sean confiables, en caso de presentarse fallas en medio de las tareas internas, los datos deberán quedar tal y como estaban antes de iniciar la transacción.

Por lo cual, las **Propiedades ACID** (acrónimo en inglés de cada una de las propiedades que deben tener las bases de datos) son:

### **A (Atomicidad)**

Las transacciones deben considerarse atómicas, es decir, se deben considerar como una sola. Esto previene que las actualizaciones de la base de datos se hagan de



manera parcial o incompleta, esto supondría grandes dificultades, por lo que, si no se completa la acción, la transacción completa se rechaza.

### **C (Consistencia)**

Consiste en que una operación no dejará datos incoherentes o incompletos o con problemas que puedan dar resultados ilógicos o ambigüedades.

### **I (Aislamiento)**

Su funcionalidad radica en aislar los datos que hacen parte de la transacción pero que no están confirmados o aún no validados. Funcionan como una especie de datos temporales mientras se ejecuta toda la transacción. Los DBMS ofrecen varios niveles de aislamiento de transacciones según sus tipos.

### **D (Durabilidad)**

Es la propiedad para garantizar que una vez se complete la transacción, los datos se actualizan y sean permanentes y confiables.

## **1.2. Técnicas de almacenamiento de datos y consultas**

En la actualidad, se presentan diferentes modelos para la organización y de datos; en esta perspectiva se tomarán las bases de datos desde dos puntos de vista, la arquitectura en relación a la estructura de los datos y el enfoque del diseño.

Desde el enfoque de la arquitectura y estructura de datos, se refiere a las bases de datos que se pueden diferenciar entre *SQL*, quienes están basados en tablas relacionadas y *NoSQL*, cuyo arreglo no se establece por tablas propiamente hablando.

Desde el enfoque de la base de datos, se refiere al uso al que se le va a dar a los datos que se almacenan, en este sentido se pueden mencionar principalmente dos:

## **Bases de datos relacionales (*OLTP*)**

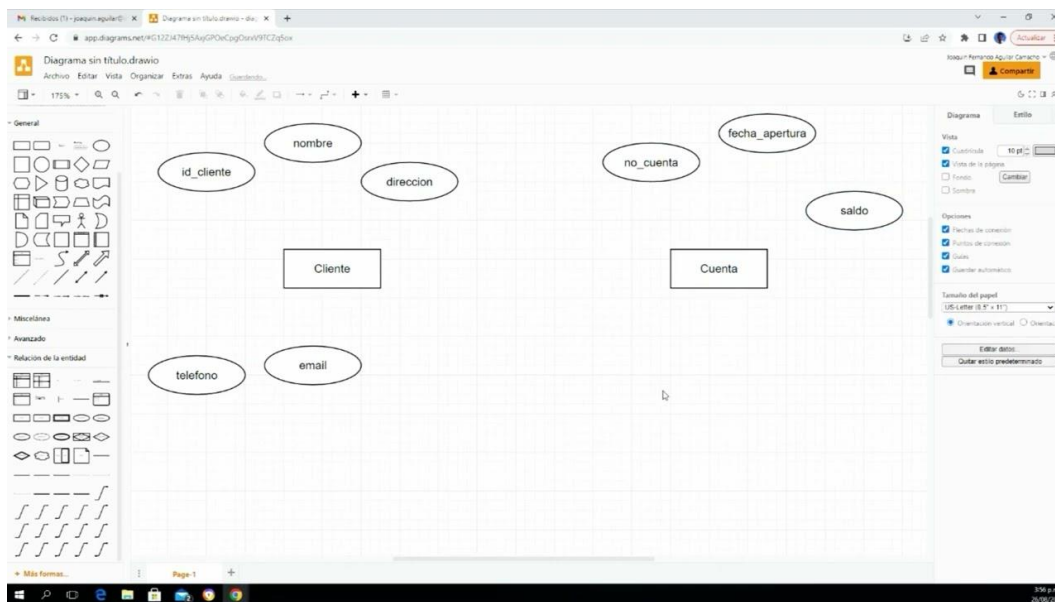
Se componen de diversas tablas que se asocian entre ellas, según las condiciones del negocio, en este sentido se determina qué entidad, mediante la definición de las cosas y actores que intervienen en el sistema, ejemplo: empleado, cliente, sede, producto, categorías, etc. Las entidades son las tablas, y cada entidad tiene atributos tales como nombre, edad, fecha de nacimiento, nombre de producto, presentación, precio, etc. Los atributos serán entonces los campos que tendrán las tablas.

Cada entidad se relaciona con otra tabla a partir de un atributo o campo en común. Este tipo de bases de datos se emplea para sistemas transaccionales o desarrollo de *software* donde son aplicaciones o desarrollos que dan solución a un proceso específico del negocio, como sistema de facturación, sistema contable, sistemas de inventarios y compras, etc. Se denomina, *OLTP* a lo que en español se nombra como Procesamiento de Transacciones En Línea.

Uno de los retos más importantes para el gestor y dinamizador de transformación digital en las organizaciones, es la identificación de estos sistemas *OLTP*, y a partir de su evaluación y preguntas del negocio determinar qué datos se requieren para iniciar su proceso hacia un sistema enfocado a la analítica.

En el siguiente video se presenta un ejemplo de construcción de un modelo entidad relación:

## Video 2. Construcción de un modelo entidad relación



[Enlace de reproducción del video](#)

### Síntesis del video: Construcción de un modelo entidad relación

Vamos a abordar una temática muy importante que es la construcción de un modelo entidad-relación. Lo primero que vamos a implementar es una herramienta desarrollada por Google llamada *draw.io online*. Para usarla, abrimos nuestro navegador y colocamos "*draw.io online*" en la barra de búsqueda. Una vez en el sitio, debemos iniciar sesión con nuestra cuenta de Gmail para crear o abrir diagramas existentes. Al elegir "crear nuevo diagrama", se nos presentan diversas categorías de diagramas que podemos construir. Seleccionamos "diagrama de entidad-relación" y procedemos a crear el diagrama.

La herramienta nos permite trabajar en la nube sin necesidad de descargar nada, solo se requiere un navegador y una cuenta de Gmail. Podemos exportar un diagrama entidad-relación desde bases de datos como *SQL Server* o *Postgres*. Para construir un modelo entidad-relación desde cero, usamos los menús de la parte izquierda, donde encontramos todos los elementos necesarios. Creamos dos entidades: "Cliente" y "Cuenta". Cada entidad debe tener atributos representados por óvalos. Para la entidad "Cliente", incluimos los atributos ID del cliente, nombre, dirección, teléfono y correo electrónico. Para la entidad "Cuenta", añadimos los atributos número de cuenta, saldo y fecha de apertura.

Es importante definir las claves principales de cada entidad. Para la entidad "Cliente", elegimos el ID del cliente como clave principal, ya que es único e irrepetible. En la entidad "Cuenta", el número de cuenta es la clave principal por las mismas razones. Una vez definidas las claves principales, relacionamos las entidades mediante líneas que apuntan a cada una, asegurándonos de que las relaciones no lleven forma de flecha. Establecemos la cardinalidad de la relación, indicando que un cliente puede tener una o muchas cuentas.

Finalmente, determinamos la ubicación de la llave foránea, que siempre va en la entidad donde está la cardinalidad de muchos, en este caso, la entidad "Cuenta". La llave foránea debe ser el mismo atributo que la clave principal de la otra entidad, en este caso, el ID del cliente. Con esto, hemos construido un modelo entidad-relación con dos entidades, cada una con sus respectivos atributos, claves principales, relaciones y cardinalidad. Los invito a consultar y estudiar cada una de las temáticas para obtener un conocimiento significativo.

## Bases de datos Dimensionales (*OLA*)

Son los diseños de datos, que están enfocados a los reportes y el conocimiento que está inmerso en los datos, este diseño permite mejorar el desempeño a los motores de bases de datos para el almacenamiento de grandes cantidades de datos. Se usan principalmente para la consolidación de bodegas de datos (*Data Warehouse - DWH*), que luego serán insumo para crear aplicaciones *OLAP* (Procesamiento Analítico en Línea) o cubos de datos.

Los cubos de datos son tablas o arreglos de datos que se componen de múltiples dimensiones, están basados en hechos, dimensiones y métricas, estos conceptos se definen a continuación:

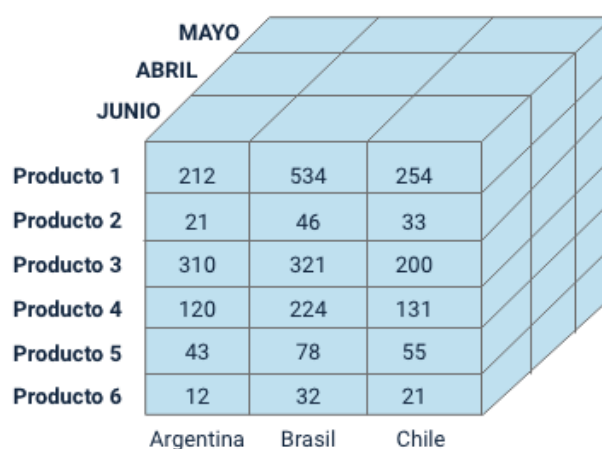
### Tablas de hechos (*Fact*)

Estas representan eventos que suceden en determinado contexto-tiempo, se caracterizan por permitir analizar los datos con el máximo detalle. Son tablas que no tienen medida y suelen ser las tablas más robustas con miles o millones de registros, además de ser 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, tablas agregadas, particionadas, etc.

| Tablas de hechos (Fact) |  |
|-------------------------|--|
| Producto_id             |  |
| Fecha_id                |  |
| Almacen_id              |  |
| Cliente_id              |  |
| Promoción_id            |  |
| -                       |  |
| -                       |  |
| -                       |  |

## Métricas

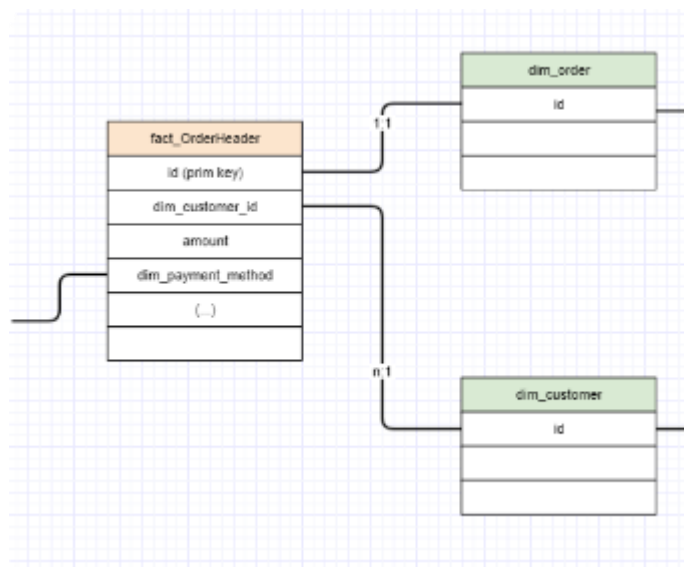
Estas medidas pueden ser conteos, por ejemplo, la participación de una persona en un evento, también se usa en las operaciones aritméticas para sumar totales, descontar impuestos, sacar medias o promedios, etc.; se utilizan para determinar contexto, es decir qué tipo de filtros aplicar o no. Suelen ser la parte más compleja, pues además de ser necesario tener el contexto de los datos y tener la modelación de lo que se desea, se debe tener manejo de lenguajes de consulta (*SQL, Python, R, Dax*, etc.).



|            |           |        |       |
|------------|-----------|--------|-------|
|            | MAYO      |        |       |
|            | ABRIL     |        |       |
|            | JUNIO     |        |       |
| Producto 1 | 212       | 534    | 254   |
| Producto 2 | 21        | 46     | 33    |
| Producto 3 | 310       | 321    | 200   |
| Producto 4 | 120       | 224    | 131   |
| Producto 5 | 43        | 78     | 55    |
| Producto 6 | 12        | 32     | 21    |
|            | Argentina | Brasil | Chile |

## Dimensiones

Igual que los hechos, son tablas; sin embargo, no suelen ser tan dinámicas como los hechos, pues las dimensiones recogen los puntos de análisis de un hecho. Por ejemplo, una venta se puede analizar respecto al día de venta, producto, cliente, vendedor, entre otros. Cada uno de estos ejemplos podrían categorizarse como dimensiones, por ejemplo: dimensión tiempo, dimensión productos, dimensión cliente, etc.



### 1.3. Estándares técnicos

Tanto para el desarrollo de *software* como para el diseño de bases de datos, se sugieren unos estándares o convenciones que, si bien no es una norma, se recomienda seguir estas buenas prácticas, en especial porque los proyectos no deben depender de personas, lo que significa que otro profesional que llegue, pueda sentirse familiarizado con los estándares y documentación.

**Algunos ejemplos de estas conversaciones para bases de datos son:**

1. Los nombres de las bases de datos deben contener el nombre de la empresa seguido del nombre de la aplicación.
  - Company\_nombreaplicacion
2. Las tablas se deben nombrar en plural, y debe estar asociado a los datos que se almacenarán.
  - Facturas, productos, detalle\_facturas

3. Por normalización, todas las tablas deben tener una llave primaria la cual debe ser id.

- Facturas -> facturas\_id. productos -> productos\_id. usuarios -> usuario\_id

De esta manera, el campo que se usa como vínculo para la relación foránea con una tabla, debe ser el nombre de dicha tabla en singular seguido del sufijo \_id.

4. Los campos de la tabla deben estar en singular y deben describir los datos almacenados.

- Usuarios (nombre, email, clave, fecha\_nacimiento...)

5. Los campos con funciones trigger (disparadores) deben estar formadas por el prefijo tgr\_ seguida del nombre de la tabla y el nombre del trigger.

- Tgr\_productos\_actualizar\_productos

Opcionalmente, los nombres pueden estar en inglés, pues la globalización y los trabajos remotos son una realidad y el inglés es el estándar mundial para la programación y desarrollo de bases de datos y aplicaciones.

#### **1.4. Ordenamiento de datos, indexación y recuperación**

La información se almacena físicamente en tablas y estas a su vez en archivos de datos que, dependiendo del diseño y motor de bases de datos, pueden influir en el desempeño y velocidad de la mismas.

Por eso es importante comprender el concepto de “**INDEXACIÓN**”. Al imagina tomar un libro, del cual se quiere leer un tema específico, ¿cuánto se tardaría buscarlo página por página?, esto sería tedioso e ineficiente. Además, se demoraría mucho tiempo, y más aún si el tema está en las últimas páginas. Lo más fácil y obvio es ir al



índice, donde se relacionan los temas y se indica el número de hoja, así se llegará más rápido. El principio de indexación de bases de datos parte, más o menos, del mismo principio del índice que se acaba de ilustrar y los administradores de bases de datos pueden gestionar estos índices en las tablas.

Un índice es un puntero o marca a una fila de una determinada tabla, es una referencia que relaciona el valor que se encuentra en una tabla con el valor determinado en el puntero. Los punteros se emplean para realizar búsquedas; en otras palabras, se pueden elegir qué columnas se hacen merecedoras de que sean buscadas para incluirles el puntero de indexación.

Esta técnica ayuda a recuperar o encontrar rápidamente los registros que se tengan de un determinado valor en alguna de las columnas. Por ejemplo, en una tabla de cliente, uno de los valores a buscar sería cédula y nombre, por lo que se le debería aplicar índices a estos campos; sino se aplican índices el valor de la cédula, por ejemplo, se le buscaría en toda la tabla, mientras si se usan los punteros se busca el registro en la columna determinada encontrando más rápidamente el valor buscado.

**Sin embargo, una tabla con muchos índices podría ser contraproducente por el tiempo de procesamiento, y además es importante usar el tipo de índice adecuado, entre los cuales se encuentran:**

**Llave primaria (*Primary key*):** cuando se configura un campo como campo clave, automáticamente se genera un índice.

- ✓ **Campos (*UNIQUE*):** aquellas columnas que se configuran como únicas, por lo general los motores de bases de datos les asignan también índice.

- ✓ **Texto completo (*Full Text*):** hay índices que se asignan a textos completos, son usados si se tiene bases de datos en los que su búsqueda es por temáticas o en los que el usuario tiene idea de qué buscar.
- ✓ **Ordinarios:** son aquellos que se puedan asignar manualmente según los criterios del diseñador de la base de datos.
- Para crear índice ordinario a través de *SQL* se emplean estos comandos:
  - *CREATE INDEX* nom\_indice *ON* table (nom\_campo);
  - *CREATE INDEX* idx\_apellido *ON* usuarios (apellidos);
- ✓ **Compuestos:** Aquellos índices que referencian a dos o más campos.
- Para crear índice compuesto a través de *SQL* se emplea estos comandos.
  - *CREATE INDEX* idx\_nombresCompleto *ON* usuarios (nombres,apellidos);

El comando *SQL* “**EXPLAIN**”, es muy usado, pues permite adelantarse a la consulta, listando las tablas en el orden que serían leídas. Es decir, presenta cómo se realizará la consulta y permite visualizar la manera en realizar la consulta, esto permite a los diseñadores de bases de datos realizar optimizaciones de ser necesario.

**Si se desea saber la manera en cómo el motor de base de datos recupera y obtiene los datos que lista, se puede realizar el siguiente ejercicio:**

- **Con la sentencia:**

*SELECT \* FROM* productos

Lista todos los campos y datos de la tabla productos.

Query 1 x concesionaria\_DBTranaccional

1 SELECT \* FROM producto

Result Grid Filter Rows: Edit:

| producto_id | marca_id | tipo_id | producto       | precio    |
|-------------|----------|---------|----------------|-----------|
| 1           | 1        | 1       | GRAND CHEROKEE | 450000.00 |
| 2           | 1        | 1       | PATRIOT        | 320000.00 |
| 3           | 1        | 1       | WRANGLER       | 380000.00 |
| 4           | 2        | 2       | FIESTA         | 250000.00 |
| 5           | 2        | 2       | FOCUS          | 280000.00 |
| 6           | 2        | 2       | MUSTANG        | 400000.00 |
| 7           | 3        | 2       | CLIO           | 250000.00 |
| 8           | 3        | 1       | SCENIC         | 300000.00 |
| 9           | 4        | 2       | BETTLER        | 320000.00 |
| 10          | 4        | 2       | JETTA          | 350000.00 |
| NULL        | NULL     | NULL    | NULL           | NULL      |

- Además, se podría aplicar una condición; por ejemplo, que muestre los vehículos que valgan más de 300.000:

Con la sentencia:

*SELECT \*FROM producto*

*WHERE precio > 300000*

Lista todos los campos que cumplen con la condición.

Query 1 x concesionaria\_DBTranaccional

1 SELECT \* FROM producto where precio > 300000

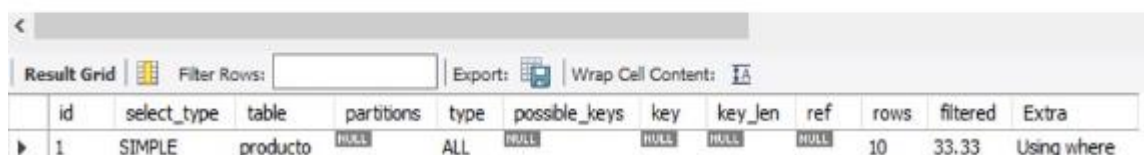
Result Grid Filter Rows: Edit:

| producto_id | marca_id | tipo_id | producto       | precio   |
|-------------|----------|---------|----------------|----------|
| 1           | 1        | 1       | GRAND CHEROKEE | 450000.0 |
| 2           | 1        | 1       | PATRIOT        | 320000.0 |
| 3           | 1        | 1       | WRANGLER       | 380000.0 |
| 6           | 2        | 2       | MUSTANG        | 400000.0 |
| 9           | 4        | 2       | BETTLER        | 320000.0 |
| 10          | 4        | 2       | JETTA          | 350000.0 |
| NULL        | NULL     | NULL    | NULL           | NULL     |

- A la consulta anterior, al adicionarle EXPLAIN al inicio, se dará un reporte de cómo se realizará la consulta:

Ejecución del comando *Explain*:

Lista que se usa un *SELECT* sencillo, que no tiene particiones de tabla, la búsqueda se realizó entre 10 campos, y que usa el extratipo *WHERE*.

| id | select_type | table    | partitions | type | possible_keys | key  | key_len | ref  | rows | filtered | Extra       |
|----|-------------|----------|------------|------|---------------|------|---------|------|------|----------|-------------|
| 1  | SIMPLE      | producto | NULL       | ALL  | NULL          | NULL | NULL    | NULL | 10   | 33.33    | Using where |

- Además de consultas y condiciones, los datos pueden ordenarse a través del comando *SQL order by*, su aplicación es muy sencilla, a la consulta anterior se le puede dar un orden, en este caso se podría tomar como elemento de orden el mismo precio. Adicional se puede definir si es ascendente *ASC* o descendente *DESC*.

Con la sentencia:

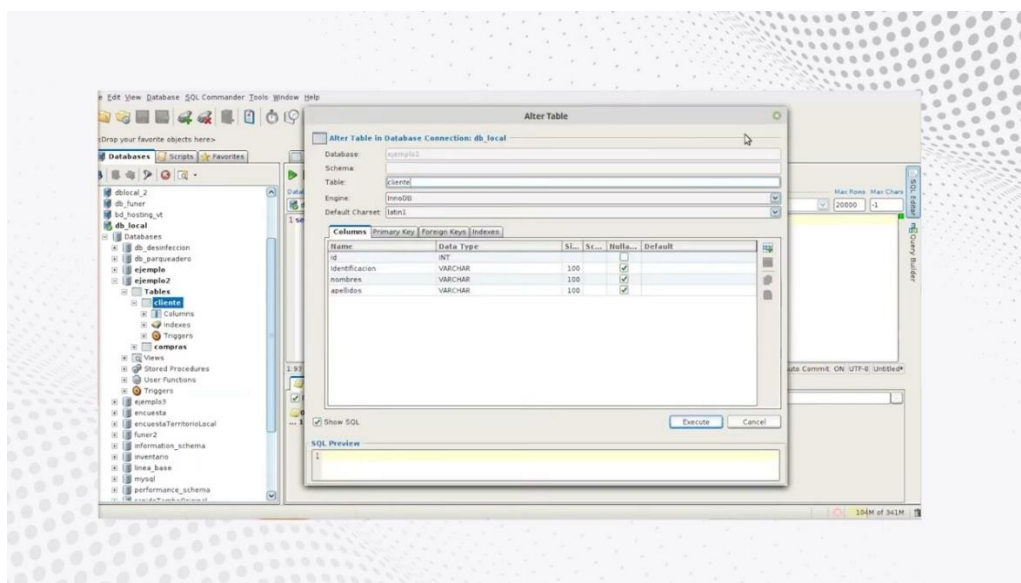
*SELECT \* FROM producto WHERE precio > 300000*

*ORDER BY precio DESC*

Lista que se ordenan los datos de manera descendente.

Para la manipulación y consulta de datos, es primordial conocer el lenguaje estándar de consulta de las bases de datos relacionales, se invita para que se amplíe el conocimiento sobre este tema en el siguiente video Ordenamiento de datos, indexación y recuperación:

### Video 3. Ordenamiento de datos, indexación y recuperación



[Enlace de reproducción del video](#)

#### Síntesis del video: Ordenamiento de datos, indexación y recuperación

Vamos a analizar el rendimiento de las consultas que realizamos en la base de datos. Aunque con pocos datos no se apreciarán grandes diferencias en los tiempos de respuesta, si se tratara de muchos usuarios o bases de datos con gran cantidad de información, la diferencia sería notable. Utilizaremos dos bases de datos similares con la misma información, pero con algunos índices diferentes. Primero, realizamos una consulta simple que cruza las tablas "Cliente" y "Compra", devolviendo todas las compras hechas por cada cliente. La consulta devuelve 146 filas en 0.003 segundos.

Luego, repetimos la consulta utilizando JOIN, obteniendo los mismos datos en 0.004 segundos. Usamos el comando *EXPLAIN* para analizar ambas consultas. *EXPLAIN* nos muestra detalles sobre cómo se ejecuta la consulta, indicando los tipos

de unión, las tablas involucradas y el uso de claves primarias y foráneas. Nos permite optimizar las consultas para mejorar los tiempos de respuesta y la calidad de la aplicación.

En general, se recomienda usar JOIN, ya que suelen ofrecer respuestas más rápidas. Para optimizar aún más, creamos un índice en la tabla "Cliente" para la identificación. Realizamos una consulta específica para un cliente, utilizando el índice recién creado, mejorando el tiempo de respuesta a 0.002 segundos.

Al utilizar índices, las consultas se optimizan significativamente. Aunque la diferencia en tiempos no es grande debido a la cantidad reducida de datos, este enfoque es esencial para bases de datos más grandes y con más tráfico. En resumen, el uso de *JOINS* y la creación de índices son prácticas clave para mejorar el rendimiento de las consultas en bases de datos, proporcionando tiempos de respuesta más rápidos y una mejor eficiencia general.

## **2. Preparación de datos**

Cuando el ser humano empezó a diferenciarse de las demás especies, una de las características más importantes, entre otras, es la manera de comportarse socialmente, es decir, cuando las relaciones sociales y comportamentales, entre manadas empezaron a cuidarse, a comunicarse, etc.

Desde ese mismo momento, la información y la manera de configurar los datos a través del proceso comunicativo (transmisor, emisor, canal, mensaje y contexto) llevan a que aparezcan diversas formas de información y datos, incluso miles de mensajes de

las primeras civilizaciones han quedado guardadas por miles de años hasta nuestros días a través de pinturas, esculturas y símbolos.

Bajo este contexto, es importante recordar la diferencia entre datos e información. Cuando las civilizaciones descubren mensajes de humanos antiguos, no hay manera de interpretar un mensaje, ante los ojos no entrenados esos mensajes solo son símbolos con valor artístico, a partir del conocimiento antropológico los mensajes representan cosas, pero realmente serían solo datos (símbolos) sin información (no interpretados).

En conclusión, los datos no son un asunto nuevo que surge, a partir de la misma evolución humana se van volviendo de igual manera complejos y bastos a medida que las civilizaciones avanzan.

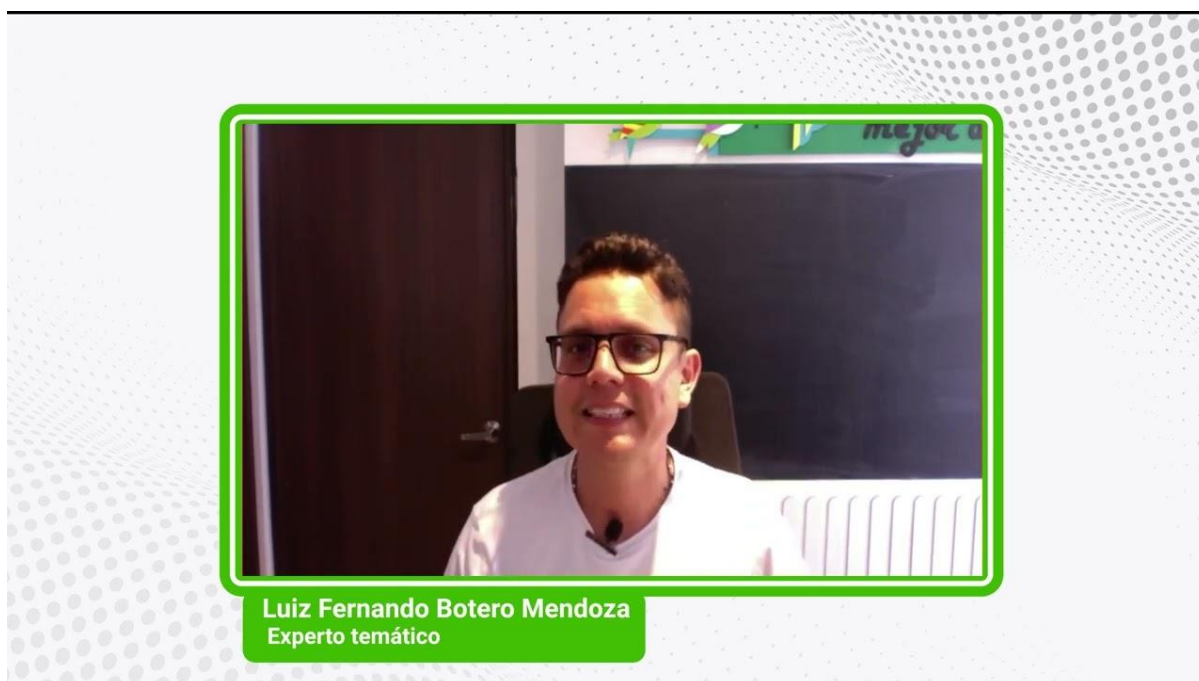
### **Datos VS Información**

Por lo que bajo el contexto digital y tecnológico actual los datos se convierten en un asunto de estudio profundo y técnico, pues el desarrollo humano y productivo tiene como plataforma principal, el uso de los datos para volverse en información, así mismo, la información deberá convertirse en conocimiento y este conocimiento es el componente principal para tomar acciones sobre la realidad.

Bajo el contexto técnico, los datos en la era moderna se guardan a través de máquinas digitales a partir de la implementación de Bases de datos (BD).

Para ampliar información se recomienda ir al siguiente video donde se profundiza en la diferencia entre datos e información:

## Video 4. Datos e información



[Enlace de reproducción del video](#)

### Síntesis del video: Datos e información

Los datos y la información son conceptos fundamentales en la era digital. Un ejemplo práctico es una *landing page* que solicita el nombre, correo electrónico y un mensaje de los usuarios. Estos elementos, conocidos como campos en programación, son datos básicos. Sin embargo, cuando se combinan estos datos para obtener un contexto significativo, se convierten en información. Por ejemplo, conocer el nombre y el correo de una persona es solo un dato, pero si además sabemos que esa persona solicita un favor o una cotización, esos datos se transforman en información valiosa. Las empresas utilizan esta información para crear estrategias de marketing personalizadas. Un supermercado que registra tus compras a través de tu número de



identificación puede saber, por ejemplo, si compras azúcar regularmente. Si dejas de comprar azúcar, pueden enviarte promociones específicas para incentivarte a volver a comprar. Así, los datos básicos se convierten en información útil y relevante. Este conocimiento permite a las empresas ofrecer servicios personalizados y estrategias de marketing más efectivas. La clave está en entender que los datos por sí solos tienen poco valor hasta que se transforman en información significativa mediante el análisis y la aplicación de estrategias adecuadas.

## 2.1. Entendimiento de la *data*

Existen varios puntos de vista de cómo mirar y analizar los datos, para empezar, los datos corresponden a los registros de transacciones o cosas que se hacen en base a un proceso de negocio. En la analítica los datos representan mucho más que la facilidad o apoyo a los procesos, significa conocimiento, evaluación, medición y mejorar las decisiones que den valor al negocio a partir de cada proceso.

Conocer su naturaleza y características es trascendental para los proyectos *BI*, donde la calidad de los datos (*Data quality - DQ*), se define como la facultad de los datos para el objetivo definido de un usuario u organización. Esto es subjetivo, ya que el concepto de calidad podría ser relativo a los estándares definidos por las expectativas de las organizaciones o usuarios (Gawande, 2020).

El objetivo final de los datos es brindar conocimiento del negocio, de una manera técnica y fiel a la realidad, en una palabra, los reportes deben ser siempre **CONFIABLES**; una unidad diminuta de información, equivalente a elementos microscópicos en medio de los océanos de información como un único registro, en uno de los campos, de

alguna tabla que componga el sistema (por ejemplo, **FECHA DE NACIMIENTO**); la colección de estos registros diversos debe tener, desde el momento de capturar los datos, una programación de validación que garantice el registro de datos tenga aspectos como: tipo de dato según la naturaleza del registro, formato uniforme aceptado por todo el sistema, que el dato sea válido por reglas de negocio y naturaleza del proceso, entre otros.

La calidad de los datos toma aún mayor fuerza para aplicar validaciones con regularidad, cuando se habla de ecosistemas de datos, donde convergen diversas infraestructuras de datos y se interrelacionan datos compartidos.

Se puede definir, que las bases de datos presentan calidad en los datos si cumplen seis dimensiones: exactitud, completitud, consistencia, unicidad, disponibilidad y validez. Sin embargo, esta clasificación no está universalmente aceptada, por lo cual se adicionan otras dimensiones: actualizado, conformidad, integridad y precisión, que complementan las dimensiones *DQ*.

**A continuación, se presenta la descripción de cada una de estas dimensiones, que son medibles y que definen la calidad de los datos o *Data Quality*:**

### **Exactitud**

Grado en que los datos representan la realidad. Por ejemplo, si un candidato para una entrevista de trabajo tiene una dirección de entrevista incorrecta, no podrá asistir hasta que obtenga la dirección correcta.

## **Compleitud**

Se establece como el porcentaje de datos poblados frente a la posibilidad de cumplimiento del 100%. Con frecuencia, se escucha, “el dato no se encuentra”. Por ejemplo, el departamento de *marketing* quiere enviar *email* a clientes, pero el funcionario de entrada de datos no completó la dirección de correo.

## **Consistencia**

Es la cercanía y uniformidad de los datos con otras tablas o un conjunto de datos de referencia. Por ejemplo, cuando en el campo “SEXO”, se espera hombre, mujer y desconocido, en algunas tablas aparece como M, F, Masculino, Femenino, *Male*, *Female*, *Unknown*, entre otras. Es importante unificar o conciliar todos los sistemas.

## **Unicidad**

Se refiere al número de veces que un objeto o evento se registra. Se espera que un hecho o entidad sea registrado una sola vez. No deberían existir datos duplicados, esto genera conteos o reportes erróneos.

## **Validez**

Se refiere a la proximidad del valor de los datos a valores predeterminados o un cálculo. Por lo general esta propiedad debe programarse o configurarse. Validez basada en reglas comerciales o cálculo (columnas o campos calculados de precios netos, total a pagar, descuentos, etc.) siguiendo las reglas de negocio.

## **Puntualidad**

Se refiere a la exactitud entre el evento real y el registrado. Existen rangos de puntualidad, algunos sistemas requieren datos en minutos, otros en horas y otros

negocios en días sin importar horas o incluso meses sin importar días (dependiendo de la granularidad).

### **Actualizado**

Se refiere a los estados del mundo real frente al estado capturado en los datos. Si los estados no se actualizan, los datos pierden calidad. Por ejemplo: Dirección cambiada: Si un cliente actualiza su dirección, los datos registrados pierden vigencia.

### **Conformidad**

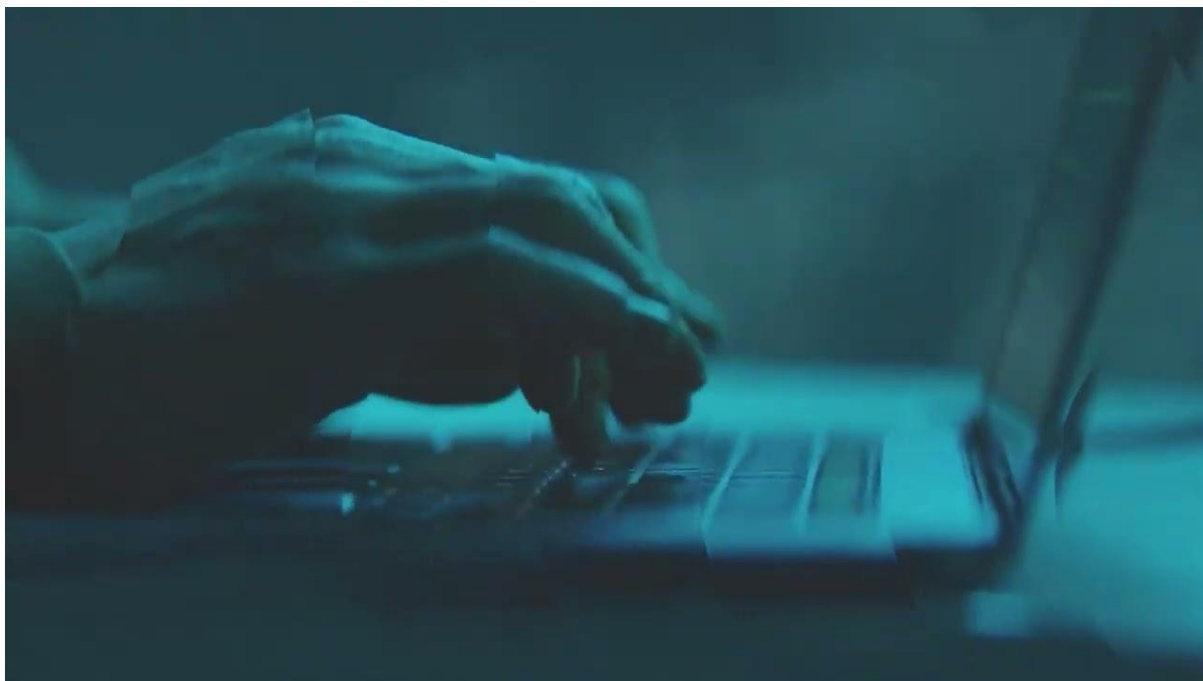
Los datos con los mismos atributos deben representarse en un formato y tipos de datos iguales. Conformidad de formato: se espera que la fecha siga el formato 'MM/DD/AAAA'. Para los humanos, los datos en la tabla parecen correctos, pero para la computadora, los cambios en el formato de los datos causarán caos.

### **Integridad**

Se refiere al grado de coherencia en que se implementa restricciones relacionales definida entre dos tablas. Integridad referencial: en la tabla padre, siempre debe existir una tabla de datos hijo. Por ejemplo: un pedido debe tener id de cliente como clave foránea.

Además, es necesario reconocer la importancia de las pruebas de integridad de datos:

**Video 5.** Entendimiento de la data - Pruebas de integridad de datos



[Enlace de reproducción del video](#)

Síntesis del video: Entendimiento de la data - Pruebas de integridad de datos

Las pruebas de integridad de datos son cruciales para asegurar la corrección y completitud de los datos en una base de datos. La integridad de datos se refiere a la precisión y consistencia de los datos almacenados cuando se realizan operaciones como inserciones, eliminaciones o actualizaciones. Esta integridad puede verse comprometida de diversas maneras, como la adición de datos no válidos, por

ejemplo, un pedido que hace referencia a un producto inexistente. También pueden producirse modificaciones incorrectas, como reasignar un vendedor a una oficina que no existe. Además, los cambios en la base de datos pueden perderse debido a errores del sistema o fallos en el suministro de energía. Otra posibilidad es que los cambios se apliquen de manera parcial, como cuando se añade un pedido sin ajustar la cantidad disponible para vender. Uno de los roles fundamentales de un sistema de gestión de bases de datos relacional (DBMS) es preservar la integridad de los datos almacenados en la mayor medida posible, asegurando que las operaciones realizadas no comprometan la consistencia y precisión de la información.

## **2.2. Detección de errores y datos faltantes**

Uno de los mayores desafíos para la analítica es lidiar con sistemas de información que fueron diseñados por fuera de los dominios de la centralización, o estándares de calidad y validación que garanticen un óptimo resultado.

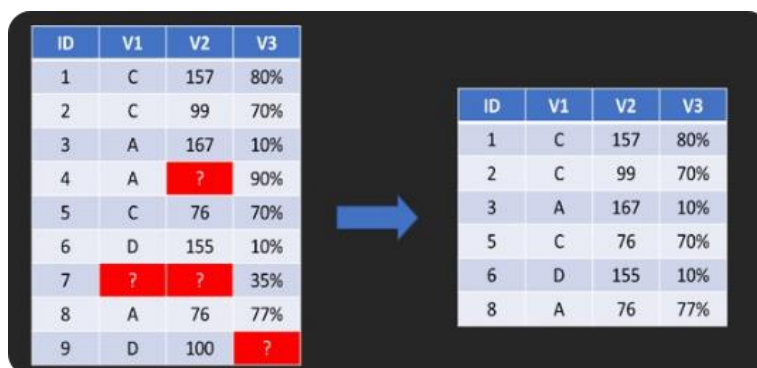
“Cuando los sistemas de información no son capaces de digitalizar eficaz y eficientemente la realidad del negocio, y capturan el nivel adecuado de detalle necesario, no almacenan dichos datos garantizando que no se producen pérdidas sintácticas ni semánticas, no procesan los datos de acuerdo con las reglas de negocio, o no listan los resultados de los análisis a los usuarios, entonces se producen “no-conformidades” en el ciclo de vida de los datos.”  
(Velthuis, 2019).

Para la exploración y la determinación de datos faltantes se debe tener presente que en gran parte del tiempo que se invierte en los procesos de analítica tiene la relación con la limpieza, depuración y mejora de la calidad de los datos y las dos maneras más comunes de gestionar los datos faltantes son: la eliminación y la imputación.

El objetivo es que los sistemas de uso analítico tengan los datos completos independiente del tipo de faltantes, y después de los análisis realizados para determinar los faltantes, y en vista de que no existe manera de calcularlos o sacarlos de otras fuentes, los datos que faltan en medio del conjunto se pueden gestionar, principalmente con las dos siguientes opciones:

### Descarte

Es la más sencilla, se puede tomar de dos maneras:




| ID | V1 | V2  | V3  |
|----|----|-----|-----|
| 1  | C  | 157 | 80% |
| 2  | C  | 99  | 70% |
| 3  | A  | 167 | 10% |
| 4  | A  | ?   | 90% |
| 5  | C  | 76  | 70% |
| 6  | D  | 155 | 10% |
| 7  | ?  | ?   | 35% |
| 8  | A  | 76  | 77% |
| 9  | D  | 100 | ?   |

| ID | V1 | V2  | V3  |
|----|----|-----|-----|
| 1  | C  | 157 | 80% |
| 2  | C  | 99  | 70% |
| 3  | A  | 167 | 10% |
| 5  | C  | 76  | 70% |
| 6  | D  | 155 | 10% |
| 8  | A  | 76  | 77% |

### Eliminación de la lista

Se remueve todo el registro de datos a los que le falta algún dato, esto tiene la desventaja que perdería otro tipo de datos, de otras columnas que sí están los datos, lo que llevaría a mayor margen de error en los reportes y dependiendo de la cantidad de datos faltantes, puede generar pérdida significativa de información.


| ID | V1 | V2  | V3  |   | ID | V1   | V2   | V3   |
|----|----|-----|-----|---|----|------|------|------|
| 1  | C  | 157 | 80% |  | 1  | C    | 157  | 80%  |
| 2  | C  | 99  | 70% |   | 2  | C    | 99   | 70%  |
| 3  | A  | 167 | 10% |   | 3  | A    | 167  | 10%  |
| 4  | A  | ?   | 90% |   | 4  | A    | null | 90%  |
| 5  | C  | 76  | 70% |   | 5  | C    | 76   | 70%  |
| 6  | D  | 155 | 10% |   | 6  | D    | 155  | 10%  |
| 7  | ?  | ?   | 35% |   | 7  | null | null | 35%  |
| 8  | A  | 76  | 77% |   | 8  | A    | 76   | 77%  |
| 9  | D  | 100 | ?   |   | 9  | D    | 100  | null |

## Eliminación por pares

A diferencia del anterior, solo pone en Nulos aquellos valores que faltan, conservando el resto de la fila, sin embargo, para modelos *BI* o *ML* podría presentar inconvenientes, pues hay medidas u operaciones que requieren la completitud de los datos.

## Imputación

Esta técnica está basada en estimar los valores faltantes en relación a los datos disponibles de la misma columna, en este caso, se hace uso de la estadística inferencial (más adelante se detalla), la idea es que los datos completados no interfieran en la media ni desviación del conjunto de datos. Este tipo de completitud se aplica a datos numéricos, ya sean tipo enteros o flotantes. Hay dos maneras de aplicar la imputación o reemplazo del dato:

| V1 | V2 | V3  |   | V1 | V2 | V3  |
|----|----|-----|---|----|----|-----|
| 25 | ?  | 50  |  | 25 | 7  | 50  |
| 27 | 3  | ?   |   | 27 | 3  | 134 |
| 29 | 5  | 110 |   | 29 | 5  | 110 |
| 31 | 7  | 140 |   | 31 | 7  | 140 |
| 33 | 9  | 170 |   | 33 | 9  | 170 |
| ?  | 11 | 200 |   | 29 | 11 | 200 |



## Imputación por media

Consiste simplemente en crear un cálculo entre los valores conocidos de la columna y ponerlo en el faltante. La desventaja es que se podrían reemplazar muchos valores faltantes con un único valor, lo que afecta la distribución de los datos.



| V1 | V2 | V3  |
|----|----|-----|
| 25 | ?  | 50  |
| 27 | 3  | ?   |
| 29 | 5  | 110 |
| 31 | 7  | 140 |
| 33 | 9  | 170 |
| ?  | 11 | 200 |

| V1 | V2 | V3  |
|----|----|-----|
| 25 | 1  | 50  |
| 27 | 3  | 80  |
| 29 | 5  | 110 |
| 31 | 7  | 140 |
| 33 | 9  | 170 |
| 35 | 11 | 200 |

## Imputación por regresión

El dato que falta se reemplaza con un valor calculado a partir de modelos de regresión, los cuales calculan el dato a imputar tomando el registro completo (todos los campos) de los datos faltantes con otros registros con datos completos, de esa manera se predicen los datos desconocidos.

### 2.3. Identificación de variables importantes

Para iniciar la toma de requerimientos para proyectos *BI*, se debe tener en cuenta que la inteligencia de negocios (*Business Intelligence*) no es una tecnología; podría denominarse más como una técnica o metodología que podría emplear una o varias herramientas tecnológicas integradas.

**Por lo cual es importante, aplicar las siguientes técnicas para la identificación de las variables:**

## **Conocer la organización (observación y lectura)**

Si es un profesional que está vinculado a la empresa o una persona que le realizará un trabajo a un cliente, lo primero es conocer bien la empresa y el contexto al cual se le realizará el proyecto de *BI*. Incluyendo el mercado, la competencia, etc.

### **Entrevista**

En la fase inicial se deben centrar en aspectos estratégicos, se debe dejar a un lado, todo lo relacionado con tecnología, es por ello, que ocasionalmente las entrevistas con los responsables de *TI* podrían ser poco fructíferas; las entrevistas de requerimientos deben estar enfocadas a las áreas que definen el rumbo del negocio, tales como la financiera, la gerencia, planeación, etc. Las preguntas deben estar enfocadas hacia las necesidades, lleva casos de uso de otras experiencias y determinar si para el área entrevistada aplican y cómo se adaptan en el caso particular.

### **Encuestas**

Actualmente las encuestas online son fáciles de crear, tienen más alcance, hay mayor control, es más cómodo para el encuestado y es más fácil de tabular. En este sentido, las preguntas, al igual que la entrevista, debe estar orientada a las variables clave y necesidad.

### **Bus dimensional**

A partir de la toma de requisitos y necesidades, se debe identificar elementos comunes. Es decir, definir elementos que en varios procesos se repiten o usan la misma información.

### **Dar prioridad a los procesos**

Se debe elegir empezar con los procesos que más impacto tienen para la organización e ir avanzando hacia las áreas y procesos menos indispensables.

### **Seleccionar las herramientas adecuadas**

Es otro elemento para garantizar la satisfacción de los usuarios finales, pues cuando las herramientas no cubren funcionalmente las expectativas, la solución *BI* podría ser poco efectiva.

### **Capacitación y cultura**

El elemento más importante en la transformación digital y la adopción con éxito de nuevas soluciones, dependen de un gran porcentaje de la capacidad y convencimiento de uso. Los usuarios deben estar comprometidos con el proyecto, las políticas corporativas deben estar enfocadas en este aspecto, es muy importante revisar las políticas, si es necesario actualizarlas hacia una nueva realidad hacerlo. Es necesario garantizar las capacitaciones y campañas hacia la cultura de los datos.

## **2.4. Dataset**

Se denomina al conjunto de datos, lo que traduce una colección de información en una sola tabla; donde cada campo representa una variable particular.

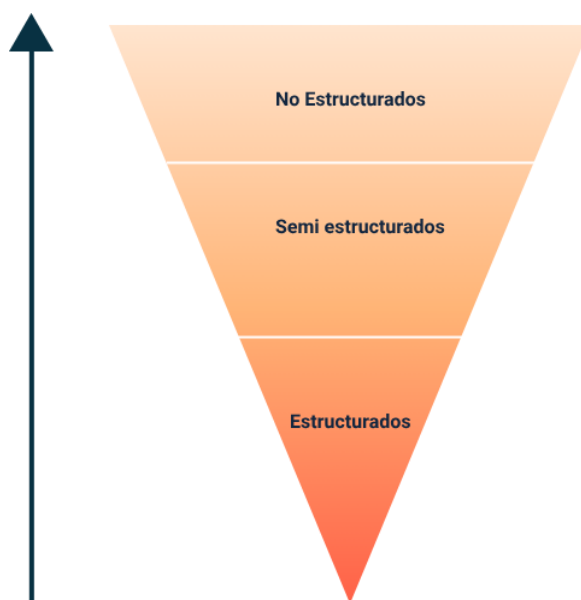
A diferencia de una base de datos donde se conforma una estructura con múltiples tablas que se relacionan, los *Dataset* contienen todo en una única tabla.

Algunas soluciones analíticas están basadas en *Datasets*, es decir, que desde los sistemas relacionales se exportan e incluyen los datos asociados de varias tablas a través de *Joins*, donde se consolidan los datos. Esta solución no es muy recomendada

en la actualidad, pues a medida que los datos se incrementan esta tabla que contiene todos los datos se hace inmanejable y poco eficiente para los motores de bases de datos.

Una de las definiciones y análisis iniciales más importantes a la hora de identificar las fuentes de datos es la manera o arquitectura en que se almacena la información. Según como se almacenan y gestionan los datos digitales, estos pueden clasificarse en términos generales como datos estructurados o semiestructurados (Omni, 2018).

**continuación se presentan los tipos de datos en los orígenes para los modelos analíticos:**



### **1. No estructurados**

No tiene un modelo de información y no está organizado en ningún formato específico. Se almacena como archivos y datos sueltos.

Ejemplos:

- Archivos de texto y videos.

- El cuerpo del mensaje de un correo electrónico.

**Más del 90% de la información disponible está en este tipo.**

## **2. Semi estructurados**

No presentan una estructura definida como en los datos estructurados, pero sí presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones, y que en algunos casos están aceptados por convención.

Ejemplos:

- **Formatos *HTML*, *XML* o *JSON*.**

- **Correos electrónicos.**

## **3. Estructurados**

Datos que se almacenan en bases de datos relacionales como *SQL*, *Oracle* etc. Los datos se organizan en filas y columnas dentro de tablas con nombres.

Ejemplo:

- **Base de datos gestión de facturación.**

La mayoría de los datos generados por todos los usuarios están clasificados como no estructurados, esto indica que se componen de información difícil de clasificar y ordenar, conformada por archivos sueltos o producidos de manera general como documentos, videos y audios.

## **3. La inteligencia de negocios**

Una característica en los últimos años en todo el mundo es la abundancia de datos que se generan día tras día, minuto tras minuto en diversas fuentes al mismo

tiempo. En las empresas esta realidad no es ajena, y toda la información corporativa generada, debería estar más organizada, segura y disponible al tiempo que se incrementa esta información momento tras momento.

Los avances tecnológicos y nuevas posibilidades de la cuarta revolución industrial hacen que este fenómeno se acentúe aún más y las empresas tengan dentro de sus retos la adopción de tecnologías y nuevas maneras de darle valor a los negocios a partir de recursos que generados por sí mismo y otros disponibles de manera pública, todo lo cual puede traducirse en oportunidades y posibilidades gigantes que bien estructuradas podría significar las mejores decisiones y valor a la organización que adopta la tecnología y las nuevas tendencias comerciales.

En forma simple, la inteligencia de negocios es una serie de técnicas, metodologías y herramientas que se integran para convertir los datos en información, luego en conocimiento para al final tomar las mejores decisiones.

### **3.1. Identificación de las preguntas básicas**

En ocasiones, los proyectos de *BI* se inician con la motivación de entrar en una ola digital, o estar en la moda de la tecnología, incluso sin tener requisitos o necesidades visibles o al menos conscientes. En caso de que las empresas, especialmente las pymes, no tengan claridad sobre el uso del *BI* en el negocio, es necesario concienciar a los directivos de que muchos problemas de la organización se deben a la falta de datos instantáneos, actualizados y que reflejan una realidad.

Una de las condiciones iniciales más importantes son las preguntas, es decir, qué se necesita saber del negocio, qué decisiones se planea tomar y qué insumos se requieren para la toma de estas decisiones.

Otro concepto errado frecuente, en especial en las pymes, es que el indicador más importante y en algunos casos el único, es el dinero que ingresa a la empresa, si bien en los negocios se trata de ganar, existen muchas variables e indicadores adicionales a los ingresos que permiten a las organizaciones avanzar mejor, aumentar su valor a partir de técnicas y conocimiento, y tras todos estos indicadores claves de negocio, deberían reflejarse en el cumplimiento de metas de ingresos y crecimiento del negocio en el tiempo.

Los indicadores y la aplicabilidad de la analítica de datos en las organizaciones deben manejarse bajo necesidades reales, preguntas fundamentales y la identificación de qué respuestas requiere el negocio para optimizar los tiempos y asertividad de la toma de decisiones.

### 3.2. Metodología de integración

La inteligencia de negocios en sí es una integración de toda la información de la organización, o al menos la más relevante. Tomar diversas fuentes (variabilidad), con información acumulada (volumen) y que además se procesen de manera óptima y rápida (velocidad).



Si bien en los libros de consulta y de manera estándar, estas tres V pertenecen al *Big Data* de manera estricta, también es aplicable a la inteligencia de negocios.

Lo primero que se debe evaluar es la integración entre las estrategias de la organización con la implementación de la inteligencia de negocio, como referencia estratégica, cada vez es más importante como apoyo a la gestión y toma de decisiones.

La analítica proporciona a las empresas capacidades, tales como ayudar a coordinar proyectos, y horarios, optimizar la asignación de personal, recursos, proporcionar la hoja de ruta para alinear con la estrategia corporativa. **El BI cambia datos internos y externos en un formato apropiado proporcionando conocimiento que apoya la toma de decisiones de los procesos o del negocio en general.**

**En la siguiente ilustración se presentan los aspectos a considerar sobre como BI integra todas o varias áreas del negocio:**





### 3.3. Herramientas de administración

Las herramientas disponibles para la implementación de la inteligencia de negocio en las organizaciones son amplias y cada vez nuevas marcas y técnicas se disputan el mercado.

**Elegir la herramienta adecuada es una variable crítica de éxito, existen múltiples aspectos qué evaluar al momento de decidirse por cuál herramienta emplear. Entre otras razones, se debe hacer análisis de:**

- Conexión a todas las fuentes de datos que la organización tiene.
- Capacidad de almacenamiento y proceso según el tamaño de los datos y la proyección.
- Relación costo-beneficio.
- **Versatilidad:** qué soporta diversas plataformas tanto como sistemas operativos como *hardware*.
- Tendencias y nuevas herramientas empleadas.

**A continuación, algunas alternativas disponibles para que las organizaciones puedan emplear con eficiencia proyectos de estas características:**

#### 6. *SAS INSTITUT*

Empresa multinacional con sede en Carolina del Norte, Estados Unidos. Es uno de los principales fabricantes de *software* de apoyo empresarial. Fundación 1976.

[https://www.sas.com/es\\_co/home.html](https://www.sas.com/es_co/home.html)

## 7. MICROSOFT

Con sus herramientas de infraestructura en la nube representada por *Azure* y la introducción de visualizaciones de *Power BI* dentro de *Power Apps* y *Dynamics 365* aumenta las capacidades analíticas integradas de Microsoft *Power Platform*.

<https://www.microsoft.com/es-mx/microsoft-cloud>

## 8. ORACLE

*Oracle Analytics Cloud* se ha convertido en la herramienta analítica y de informes imprescindible para las pequeñas empresas, más allá de las grandes empresas a las que suele servir.

<https://www.oracle.com/lad/business-analytics/analytics-services>

## 9. INFORMÁTICA

Integrada con *Azure*, cuenta con su herramienta *Cloud Data Marketplace*, diseñada para aumentar de forma intuitiva el intercambio de datos, mejorar la productividad y permitir que las organizaciones tomen decisiones más informadas.

<https://www.informatica.com/products/cloud-data-integration.html>

## 10. QLIK

Empresa especializada en analítica y procesos ETL, Herramientas especializadas de desde punto inicial al punto final.

<https://www.qlik.com/es-es/products>

Existen otras más, soportadas con grandes marcas como *IBM, Google, Teradata, SAP* y otras que si bien no son tan mencionadas en el mercado, son igual de poderosas y confiables.

### **3.4. Técnicas de solución de problemas (modelación de datos)**

Si bien existen varias metodologías, todas coinciden en una manera genérica para la solución de un proyecto analítico propuesto por Davenport, con algunas variables.

**Esta metodología tiene tres grandes etapas:**

#### **Definición del problema**

Para reconocer el problema, se debe partir de las necesidades surgidas a partir de la experiencia para tomar una decisión o acción; es decir, qué necesidades o inconvenientes se han tenido cuando la información, reportes o conocimiento de algún dato importante no está disponible en el momento y exactitud para tomar decisiones o ejecutar acciones. Es importante gestionar las expectativas y documentar las necesidades o problemas puntuales.

**En esta fase, una clave es identificar objetivos que no se cumplen por la falta de un proyecto analítico.**

#### **Resolviendo el problema**

Después de la primera etapa, donde se tiene claridad de los problemas, el paso siguiente es buscar solución. Por lo general está compuesta de cuatro etapas:

- Modelar y seleccionar variables que representan el problema y determinar las variables que forman parte de él.

- Recopilar datos para identificar dónde se consiguen datos para las variables del modelo.
- Análisis de datos. Determinar naturaleza, fuentes y gestión para el acceso técnico de la información.
- Desarrollar componentes técnicos y generación de reportes con información de valor y conocimiento del negocio.

### **Comunicación y actuación en función de los datos**

Cuando se generan procesos de analítica, es tan importante, la comunicación y uso de los datos para el negocio como su desarrollo; el paso final de estos proyectos es comunicar y accionar los resultados, es necesario contar una historia a partir de los datos y a partir de ello generar conocimiento para actuar mejor.

La modelación de las soluciones *BI* depende de las reglas de negocio y los requisitos o problemas a resolver. En los apartados siguientes se detallan las arquitecturas disponibles y el proceso del tratamiento de datos a partir de una mirada de análisis e inteligencia de negocio.

### **3.5. Metodologías de análisis (*Kimball*, *Inmon*)**

En lo relacionado a la planeación y diseño de las bodegas de datos (*DWH*), se plantean dos arquitecturas:

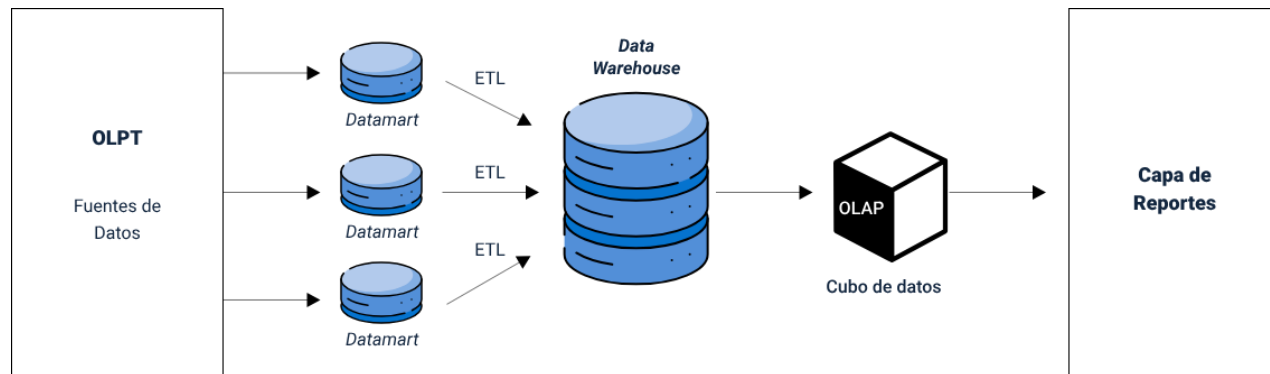
#### **Modelo *Kimball***

La metodología para el diseño de *DWH* propuesta por Ralph Kimball se centra en copiar las bases de datos transaccionales en un modelo optimizado para consultas de analítica. Para integrar varias áreas o disciplinas del negocio se emplean *data marts*,

que son bases de datos que surgen a partir del proceso de transformación del ETL por cada área o división del negocio.

**Kimball propone que estos *Data Mart* deben crearse primero para proporcionar capacidades analíticas;** en pasos siguientes del flujo de datos se integran en una bodega de datos empresarial de manera integral. Una desventaja de este modelo es que al cambiar los *Data Marts* se podrían perder dimensiones en los reportes diseñados. Adicionalmente, se podrían presentar datos redundantes en varios *Data Mart*. No obstante, es el modelo más usado en las soluciones *BI*, no indicando esto que sea mejor (consultar figura).

**Figura 1. Modelo Kimball**

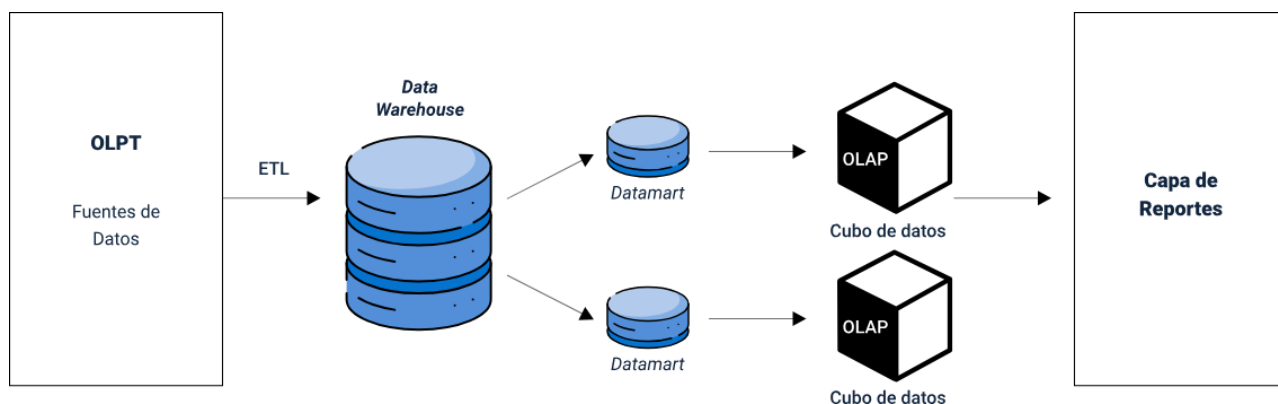


### Modelo Inmon

No es muy diferente a Kimball, los principios de transformación y carga a los *DWH* contienen los mismos elementos en su arquitectura, sin embargo, la metodología de Bill Inmon plantea que los *data marts* deben ir después del almacenamiento de *DWH*.

La metodología Inmon comienza por generar un arreglo de todos los datos corporativos en el *DWH*, para luego identificar y dividir áreas para la generación de los *Data Mart*. En este modelo se normalizan más los datos, conteniendo menos redundancia; es un poco más complejo de usar a nivel comercial o abierto, sin embargo, podría ser más administrado en el sentido que se crean los *data marts* y cada área del negocio podría usarlos creando sus propios cubos de datos o reportes (explorar figura).

**Figura 2. Modelo Inmon**



### 3.6. Verificación de valores y escalas

Los cuadros de mando o *Dashboard*, permiten hacer monitoreo de los procesos mostrando información importante a través de elementos gráficos de fácil entendimiento y con actualizaciones periódicas.

Los valores que presentan deben ser indicadores de procesos, tareas o situaciones importantes para el negocio, su escalabilidad y visualización deben aparecer

claramente. Para algunos valores podrían presentarse escalas (ejemplo, valores de ventas), para otros no (ejemplo, comparaciones porcentuales).

**Entre los elementos más importantes de los cuadros de mando, su tipo de valor y escalas se pueden mencionar:**

### **1. Tablas**

se conforma de matriz, presenta múltiples datos, puede ser estática o dinámica según las reglas del negocio y características de la información que representa. Presenta datos estructurados por dimensiones (tipo *OLAP*) en algunos casos, y su escalabilidad se limita al redondeo de cifras configurando la cantidad de decimales de los datos numéricos flotantes.

### **2. Métricas**

los valores surgen como resultado de una actividad específica y las medidas son el resultado de estas actividades en su conjunto o segmentación, siguiendo o no una serie de condiciones y operaciones. Por lo general las métricas se denominan KPI.

### **3. Listas**

comúnmente formadas por KPI. En caso de que el cuadro de mando solo esté formado por este tipo de elemento, se denomina *Scorecard*.

### **4. Gráficos**

el fin es mostrar datos con alto impacto visual, que sirva para obtener información acumulativa o calculada. Al igual que las tablas pueden tener múltiples dimensiones, pero se verán con mejor presentación y dimensión los datos. Se debe poner especial atención en el aspecto de la escalabilidad, pues hay datos que por su

tamaño deben ser escalados para que se compare y se note las diferencias; sin embargo, ocasionalmente no es siempre recomendable, por ejemplo: un gráfico de barras si presenta la cantidad de dinero vendido, es factible escalar, para que el valor no empiece en 0 (cero), sino por ejemplo en 10 millones (según el tamaño de las cifras), pero si estas barras presentan porcentajes, es necesario que la gráfica siempre empiece desde el 0%, para que se note bien la dimensionalidad de los datos.

Este es el elemento más común y de mayor variabilidad, existen múltiples opciones de visualizadores gráficos en diferentes segmentos (barras, circulares, mayas, etc.).

## **5. Mapas**

este elemento permite mostrar información geolocalizada. Aplica a los datos de ubicación, de esta manera se dimensiona la ubicación de las cifras que se estén representando.

## **6. Alertas visuales**

por lo general, al momento de desarrollar los cuadros de mando se pueden incluir a la programación de alertas automáticas con el fin de informar un acontecimiento crítico, ya sea por fechas u otro evento que suceda.

## **7. Menús de navegación**

estos elementos visuales, ya sean en texto o botones, facilitan al usuario navegar y realizar operaciones interactivas en los elementos del cuadro de mando.

Ejemplo de componentes de un cuadro de mando o *Dashboard*:





### 3.7. Procedimientos almacenados y funciones

Desde los motores de bases de datos, se pueden aplicar funciones nativas a los datos, es decir, no todas las condiciones y reglas de negocio se programan en los procesos de programación, pues los motores de bases de datos pueden adaptar además de las consultas estándar *SQL*, *script* con lenguajes como: R, Python, etc.

Desde el diseño de los datos y la analítica se pueden crear y usar una cantidad de funciones, amplias en cantidad y diversidad de usos, entre las más comunes, los expertos en datos aplican consultas de todo tipo empleando condiciones, filtros y contexto de datos; adicionalmente pueden crear medidas que calculen o cuenten elementos de datos, filas calculadas a partir de datos de otras filas o funciones.

### 3.8. Disparadores

Llamados comúnmente *Triggers*, son sentencias o funciones que se ejecutan cuando se presentan ciertos eventos. Estos eventos pueden ser una condición en el tiempo, una actualización, inserción o borrado de una tabla en un dato específico.

Los disparadores son elementos muy importantes a la hora de automatizar acciones en los datos. La tendencia en los sistemas de información es la automatización y dejar todo lo posible para que las máquinas funcionen de manera autónoma, de esta manera se tendrán acciones en los datos en el momento mismo de realizar una operación que sea el disparador de otras tareas (explorar tabla).

**Tabla 1.** Ejemplo de *Triggers*

| Disparador   | Acción  |
|--|---|
| Cuando en la tabla clientes se realiza un <i>insert</i> (cliente nuevo). | Enviar correo electrónico con plantilla de bienvenida y resumen de los datos registrados.           |
| Todos los días cuando el sean las 3:00 A.M.                              | Ejecutar carga de la base de datos transaccional a los almacenamientos para el proceso <i>ETL</i> . |

En esta tabla se presentan ejemplos de disparadores, para una mejor comprensión.

#### 4. Análisis exploratorio de datos

La analítica basa sus procesos en conceptos estadísticos apoyados de herramientas tecnológicas, tanto para el procesamiento de los datos (muchos datos de entrada, procesados en el menor tiempo posible) como para la presentación o visualización de los mismos, en este apartado se señalan algunos conceptos teóricos, fundamentados en la matemática para luego aplicar analítica de datos.

#### **4.1. Estadística descriptiva y estadística inferencial**

Son los métodos mediante los cuales se presenta la información y pueden clasificarse en:

##### **Estadística descriptiva:**

son los datos que se pueden representar a partir de tablas, gráficos y otros recursos. Describe fenómenos, por ejemplo, cuando se pregunta la edad a un grupo de personas, se podría realizar tabla o gráfico con estos resultados y definir medidas descriptivas como edad promedio, edades más frecuentes, etc.

La analítica usa principalmente este tipo de estadística, pues parte de que no se tienen conocimientos previos ni supuestos verdaderos, describe fenómenos que pasan en el negocio, además, por lo general se tienen todos los datos disponibles para los reportes, por lo que pocas veces se emplean muestras poblacionales.

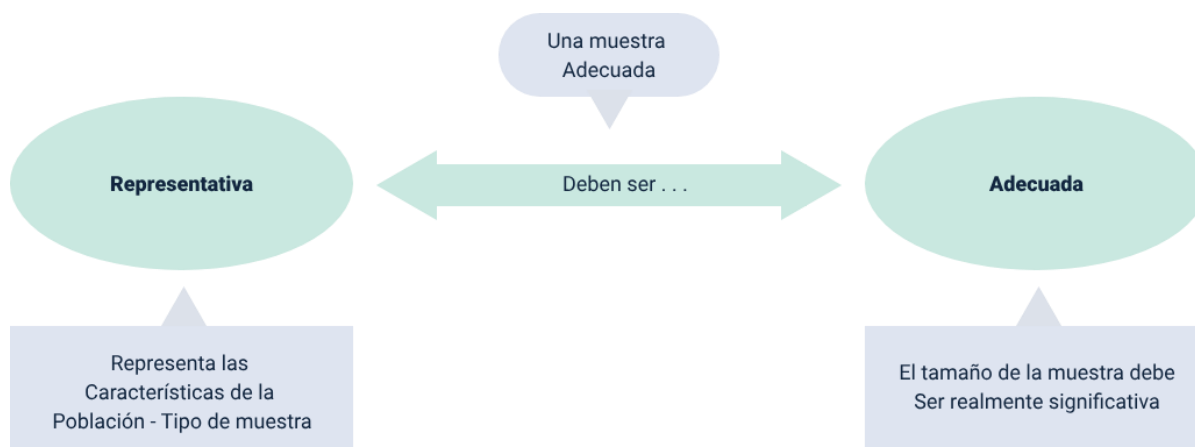
##### **Inferencia estadística:**

a partir de métodos, se pueden realizar conclusiones, tomar decisiones, estimaciones o predicciones sobre una población o universo, con base a datos de una muestra. Para la aplicación de estadística inferencial, se tratarán algunos parámetros matemáticos para la aplicación de este modelo estadístico.

#### **4.2. Población y muestra**

Para la aplicación de estadística inferencial, se tratarán algunos parámetros matemáticos para la aplicación de este modelo estadístico que se ilustran a continuación:

**Figura 3.** Características de una muestra adecuada



Este tipo de estadística se aplica cuando, la población o el universo objeto de estudio es muy grande e imposible de aplicar estadística descriptiva o aplicar instrumento a todo, o cuando no hay exactitud del número de población o universo. Para estos casos, se deberá aplicar la estadística inferencial donde se pueden realizar afirmaciones sobre una población basado en los resultados de una muestra.

**Es importante** no establecer divorcios entre la estadística descriptiva y la inferencial, ambas son necesarias, pues la inferencial usa los datos descriptivos para llevar a conclusiones generales.

#### 4.3. Escalas de medida y clasificación de variable

Las escalas de medición son procesos de comparación y dimensionalidad, que da cuenta de un valor que signifique sus proporciones. Una variable, es un elemento que cambia y que al hacer parte de un conjunto puede afectar.

Las escalas se convierten en algo real a través de las preguntas que se utilizan para recolectar la información aplicando la escala. (Domínguez, 2017). Existen cuatro escalas primarias de medición: nominal, ordinal, intervalo y radio; algunas escalas más

sofisticadas como las escalas multicontenido o multipropósito, en la siguiente tabla se describen las categorías.

**Tabla 2.** Variables y escalas

| Variables | Escala    | Descripción  |
|-----------|-----------|--|
| Categoría | Nominal   | Clasifica los elementos del conjunto para distribuirlos en grupos.   |
|           | Ordinal   | También clasifica elementos, pero además permite hacer escalas de medición comparativa.  |
| Métrica   | Intervalo | <p>Cuantifica y califica numéricamente los objetos de la categoría.</p> <p>Permite hacer mediciones simples.</p> <p>Utiliza escalas continuas.</p> <p>No tiene cero por lo que usualmente aplica escalas pares; la más común es de 1 a 10. Cuando la escala no tiene un punto neutral, como si lo tiene una escala impar, el investigador estaría forzando una respuesta negativa o positiva del participante. De 1 a 5 es negativa y de 6 a 10 es positiva.</p> |
|           | Radio     | <p>Cuantifica y califica numéricamente los objetos de la categoría.</p> <p>Permite hacer mediciones simples.</p> <p>Utiliza escalas discontinuas.</p> <p>Resuelve el problema del cero por lo que usualmente aplica escalas impares; las más comunes son de 1 a 5 o de 1 a 7.</p>  |

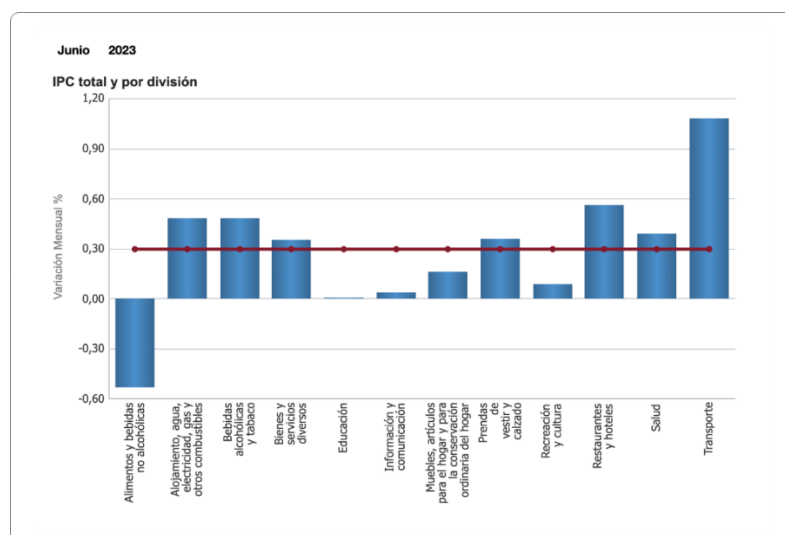
#### 4.4. Técnicas de análisis estadístico

Mencionar parte de las técnicas estadísticas sería un muy extenso y no bastaría un fragmento de un recurso de formación, sin embargo, para efectos de contexto y orientación hacia la analítica de datos y soluciones *BI*, es fundamental señalar que el análisis estadístico se asocia con las técnicas y prácticas propias del *Big data* y la inteligencia de negocios.

Pero desde el enfoque propiamente, la utilidad estadística está inmersa como fundamento en el conocimiento de las áreas y en todas las decisiones incluso las decisiones no técnicas y simples, de manera implícita se trae a la mente procesos estadísticos (ejemplo: al pasar una calle, debes verificar variables como flujo de carros, velocidad, estado de la vía, distancia del punto A al punto B, etc.).

Para todo negocio, es importante tener en cuenta datos financieros y de consumo (consultar figura).

**Figura 4.** Índice de precios del consumidor



Nota. Banco de la República (2022).

En la imagen anterior se presentan los índices de precios, en los que se nota la inflación y consumo de la población en Colombia. Se nota que las bebidas alcohólicas y tabaco tuvieron estabilidad de precios a lo largo del mes de julio. Siendo el calzado y prendas de vestir las que más tuvieron variación en sus precios, seguido de los alimentos.

Como principio fundamental la estadística contiene los siguientes componentes:

### **Objetivo y preguntas**

Para empezar a hacer uso de la estadística, es fundamental definir la intencionalidad de lo que se desea saber, comprobar o medir. En otras palabras, definir con claridad las preguntas que se desean responder.

### **Recopilación de los datos**

Según la necesidad u objetivo del ejercicio estadístico, es necesario identificar y si no existe, diseñar la herramienta o instrumento para la recopilación de los datos. Verificar si los sistemas de información tienen los datos requeridos, implementar encuestas, o métodos de observación sistemática.

### **Procesar los datos**

Cuando se tengan los datos recopilados de manera sistémica, es importante tener claridad de qué hacer con ellos. Tener los datos y ya ocasionalmente no responde las preguntas requeridas si el proceso no está claro. Es fundamental interpretar y entender los datos para definir las fórmulas matemáticas y las mediciones para aplicar las operaciones a estos datos. Las operaciones estadísticas más comunes y simples son: Promedios, sumas, conteos, segmentación, porcentajes, variables, tiempo, etc.

## **Analizar los datos**

Este análisis, el cual se presenta junto a la presentación de datos, está basado en la interpretación y comprensión de la información que hay en los datos presentados. Se definen también con medidas como desviaciones, comprobaciones estadísticas, proyecciones futuras, entre otras operaciones para determinar el estado de lo que se ha trazado desde los objetivos y preguntas.

## **Conclusión y acciones**

Si bien la estadística no se centra en las acciones, el fin es dar un conocimiento del negocio para que se use como insumo en la toma acertada de decisiones.

La estadística es entonces, un conjunto de métodos y teorías aplicadas a la recolección, descripción y análisis de datos, los cuales constituyen evidencia numérica para la toma de decisiones en condiciones de incertidumbre.

## **5. Métodos para hacer análisis exploratorio de datos**

La exploración de datos es sin duda, una de las tareas más importantes, pero también es una de las más complejas, especialmente para iniciar su exploración. La primera tarea es tener idea de la manera como están los datos, identificar las variables más importantes, la manera en cómo se relacionan o no unas con otras, tamaño de los datos, determinar si se presentan patrones, cálculos, qué calidad tienen estos datos, etc.

Puede ser una tarea dispendiosa pero necesaria antes de emprender procesos técnicos de inteligencia de negocios.



### 5.1. Datos univariantes

Son los datos que solo se ocupan de una sola variable, por lo tanto, este tipo de datos son la forma más sencilla de analizar, pues la información se ocupa de una sola cantidad que cambia de registro a registro. Describe los datos y encuentra los patrones que se puedan presentar.

Por ejemplo: la altura de las personas. En este caso la única variable es la altura y su análisis se centra en determinar la media, el máximo, mínimo, moda, etc.

### 5.2. Datos bivariados

Son aquellos datos que involucran dos variables. Su análisis toma causas y resultados asociados para conocer la relación entre las dos variables.

**Un ejemplo** de datos bivariados puede ser la temperatura del ambiente en un día y las ventas de helados.

### 5.3. Datos multivariantes

Similar al anterior (bivariado) pero contiene más de dos variables dependientes.

**Su análisis suele ser más robusto según los objetivos planteados del análisis de este tipo de datos, se puede emplear análisis de regresión, el análisis de ruta, el análisis factorial, etc.**

### 5.4. Reglas de negocio

Desde el punto de vista de programación, el cual debe coincidir con el negocio, son las condiciones asociadas a las tareas que forman los procesos. Las reglas no son la normatividad de las empresas, se refiere más a los requisitos o necesidades y cómo los sistemas de información satisfacen estos requisitos bajo las reglas y condiciones técnicas necesarias.

**Las reglas de negocio se basan en las políticas de las organizaciones, en metodologías ágiles como el Scrum, las reglas de negocio se plasman en historias de usuarios.**

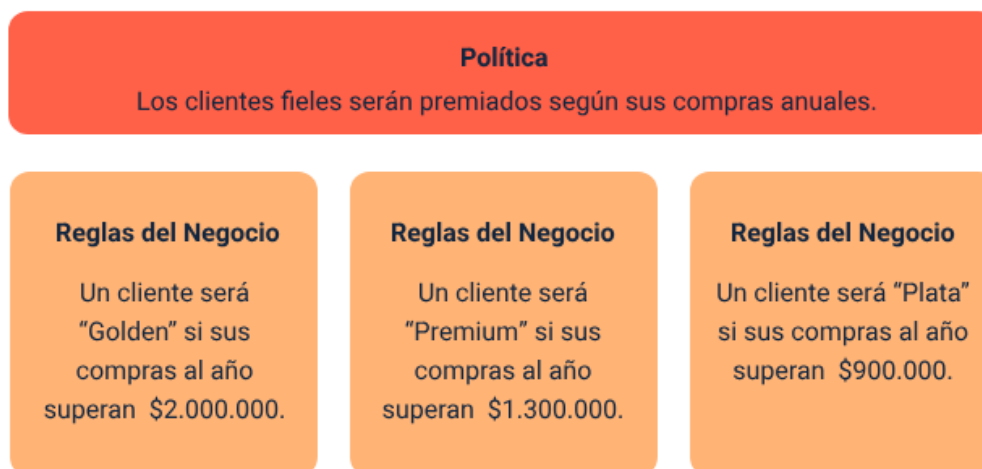
Por ejemplo:

- ✓ Como cliente.
- ✓ Quiero pagar los artículos del carrito de compras.
- ✓ Para recibirlos en casa.

En medio de este requerimiento, es posible que existan políticas de la organización que restrinjan la acción solicitada, pues podrían existir otras reglas que indiquen (figura 5).

- ✓ Como propietario de la tienda online.
- ✓ Quiero rechazar compras fuera de Colombia con pedidos inferiores a \$300.000 pesos.
- ✓ Para evitar gastos de envío que hacen la compra no rentable.

**Figura 5.** Características de una muestra adecuada



A estas reglas de negocio, a menudo se les denomina también Lógica de negocio.

### 5.5. Tipo de restricciones

Las restricciones limitan las acciones que el sistema o los usuarios pueden realizar. Las reglas por lo general son restrictivas, es decir, se debe o no se debe hacer algo a no ser que cumpla condiciones como roles de usuarios, valores previos, etc. Las restricciones, no deben convertirse en inconvenientes al usuario o sistema y los mensajes que se generan deben ser muy claros.

Existen diferentes tipos, que se detallan a continuación:

#### Validación de datos

Son aquellas que vienen como norma del tipo y formato de datos, esto se hace para garantizar la calidad y completitud de la información.

## **Políticas de la organización**

Son aquellas que surgen a partir de las reglas de negocio y su naturaleza surge a partir del funcionamiento de los procesos. Por lo general se programan en procedimientos o algoritmos programados que evalúan las condiciones y restricciones previas.

## **Normatividad regulatoria**

Muchas condiciones provienen de regulaciones del gobierno, según el sector del negocio y las regulaciones que por ley existen. Un ejemplo podrían ser los sistemas de formación y certificación de los centros de entrenamiento para certificar trabajo seguro en alturas a trabajadores, entre otras restricciones un instructor no podrá tener más de 35 alumnos por cada grupo.

### **5.6. Programación transaccional**

Su característica principal en las organizaciones, son las aplicaciones que emplean bases de datos, estas pueden ser *SQL o NoSQL*, pero indistintamente, son datos que se asocian y presentan dependencia entre sí.

Los datos que surgen de estos sistemas son fuentes para los procesos analíticos, los cuales toman varios programas transaccionales.

### **5.7. Programación de estructuras no lineales, desnormalización, series y *dataframes***

En una estructura lineal, cada elemento sólo puede ir enlazado al siguiente o al anterior. A las estructuras de datos no lineales se les llama también estructuras de datos multienlazadas y tiene las siguientes características:

✓ **Enlazado**

Cada elemento puede estar enlazado a cualquier otro componente.

✓ **Sucesores**

Se trata de estructuras de datos en las que cada elemento puede tener varios sucesores y/o varios predecesores.

✓ **Estructura**

Su aplicación se hace en estructuras de árbol o grafos.

## **5.8. Álgebra relacional**

Es el área del álgebra que usa métodos para crear nuevas relaciones a partir de unas ya existentes. Todas las operaciones sobre tablas relacionales a través de un lenguaje de manipulación de datos están bajo este esquema.

Emplea operadores y otros elementos del álgebra, existen entre otros operadores de proyección, selección, unión, diferencias intersecciones, divisiones, etc.

Este tipo de operaciones son muy usadas en los procesos *ETL*.

Por eso es importante tener presente las siguientes recomendaciones entregadas en el siguiente video:

### Video 6. Álgebra relacional - Manipular los datos



Video. SENA 2024.

[Enlace de reproducción del video](#)

#### Síntesis del video: Álgebra relacional – Manipular los datos

Para manipular los datos y desarrollar un análisis de datos exploratorio efectivo, es necesario tener claridad sobre los tipos de valores que pueden presentarse durante el proceso de manipulación de datos. Para identificar estos valores nulos dentro de la colección, se debe utilizar el comando adecuado. La primera parte del comando hace referencia a la variable donde se almacenan los

registros del archivo CSV. La segunda parte del comando determina cuáles de esos registros en cada una de las columnas están vacíos, y la tercera parte permite determinar la suma del total de los registros nulos por cada una de las columnas. Como se puede apreciar en la siguiente figura, se identificaron dos variables con valores nulos para un total de cuatro registros. Al finalizar el análisis de los resultados obtenidos y tabular la información, quedaría de la siguiente manera: la decisión de eliminar los registros nulos depende de cada analista de datos, quien debe considerar si estos valores pueden afectar significativamente el análisis. Es importante revisar la cantidad de datos a eliminar; en el caso del ejemplo, solo se deberían eliminar cuatro registros, lo que no tendrá mayor afectación en una muestra de 1349 datos.

Si se divide el total de valores duplicados sobre el total de registros, se obtiene un porcentaje bastante bajo, por lo que se procederá a eliminar dichos registros. Para la eliminación de datos se utiliza el comando que aparece en pantalla, asignando el resultado nuevamente a la variable `df`. La primera parte recibe el resultado con los datos nulos eliminados, la segunda parte recibe la variable de los datos originales, y la última parte elimina las filas con los valores nulos de los datos originales. Finalmente, se ejecuta el comando `df.in` para revisar los resultados luego de eliminar los valores nulos. El resultado obtenido muestra que todos los valores están con el mismo número de registros, con un total de 1345.

Ahora, para eliminar los valores duplicados, es muy fácil. La duplicidad de los datos puede afectar significativamente la muestra, por lo que se debe aplicar el comando adecuado para eliminarlos. La primera parte del comando recibe el resultado con los datos duplicados eliminados, la segunda parte del comando determina la variable de datos originales sin valores nulos, y la tercera parte elimina

los registros duplicados en el conjunto de datos `df`. Nuevamente, se debe ejecutar el comando `df.in` para revisar los resultados luego de eliminar los valores duplicados. El resultado generado es el que evidencia en la siguiente imagen, y la tabulación de los datos quedaría de la siguiente manera.

Finalmente, si se tabula toda la información, teniendo en cuenta que ya se han descartado valores nulos y valores repetidos, los resultados se pueden explorar en la siguiente tabla. Para ordenar los datos, puede hacerlo con la siguiente línea de comandos; ejecútela y analice los resultados. Para el proceso de manipulación de datos, es importante conocer el proceso para el agrupamiento de datos. En este caso, se utilizará el rango de edad, ya que es una de las variables objeto de análisis, y agruparla permitirá tener mejor claridad en el procesamiento de los datos. Para realizar el procesamiento y el agrupamiento de datos, se deben seguir los siguientes pasos: definir los rangos de edades, crear una variable denominada rangos donde se colocarán los números que equivalen a cinco cortes, y establecer el nombre para cada rango. A estos rangos se les debe asignar un nombre, para lo cual se define una variable denominada `nombre_rango` y se le asignan letras a cada rango.

Para generar este rango de edades, se crea una nueva columna denominada `rango_edad` donde se almacenará cada una de las nuevas variables, dependiendo del rango al que corresponda. Para esta acción, se ejecuta la línea de comandos que se observa en la pantalla. El gráfico a continuación ejemplifica cómo quedarían agrupados los datos por rango de edades, a los cuales se les asigna una letra. El resultado que se obtiene sería el siguiente: como se presenta en la imagen, al final se ha creado una nueva columna agrupando el rango de edades correspondiente.



## 5.9. SQL

*SQL* (por sus siglas en inglés *Structured Query Language*; en español lenguaje de consulta estructurada) es un lenguaje que da acceso a un sistema de gestión de bases de datos relacionales que permite especificar diversos tipos de operaciones en ellos.

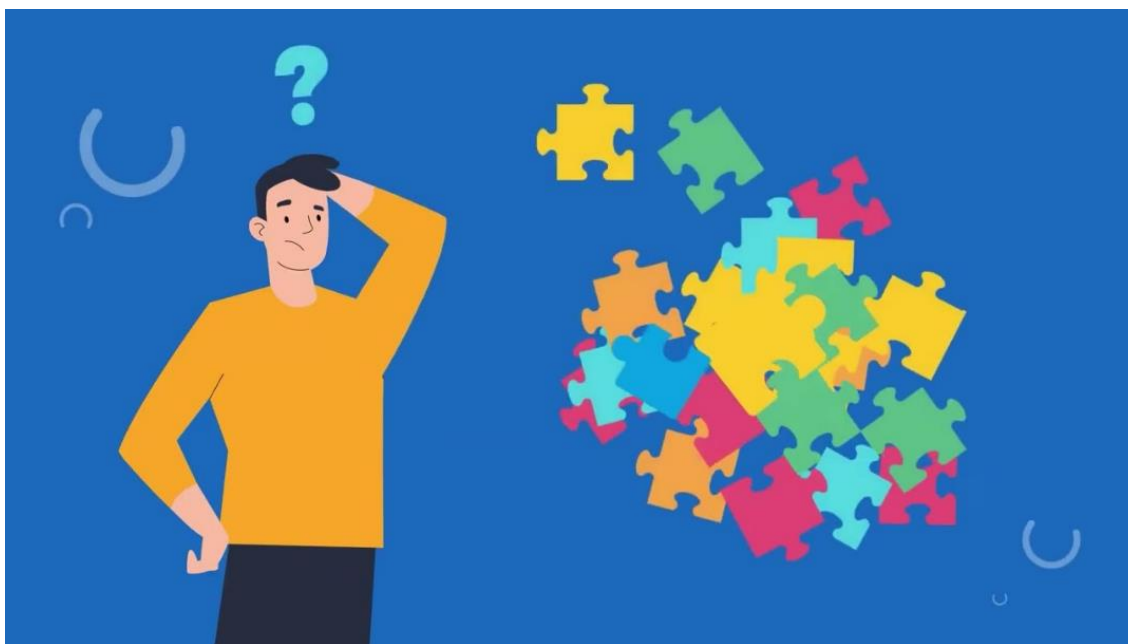
Es un lenguaje estándar consolidado por el Instituto Americano de Normas (*ANSI*) y por la Organización de Estándares Internacional (*ISO*). Está compuesto por comandos, cláusulas, operadores y funciones de agregado.

Es el lenguaje “natural” para el manejo y consulta de datos, todos los sistemas relacionales por lo general, lo emplean para la manipulación de datos complementando con otros lenguajes.

Existen varios motores de bases de datos con núcleo *SQL*, entre los más conocidos *MySQL* de uso libre y otra distribución licenciada por *Oracle*.

En el siguiente video se presentan las ventajas de *Mysql*:

### Video 7. SQL - Ventajas de MYSQL



Video. SENA 2024.

[Enlace de reproducción del video](#)

#### Síntesis del video: SQL – Ventajas de MYSQL

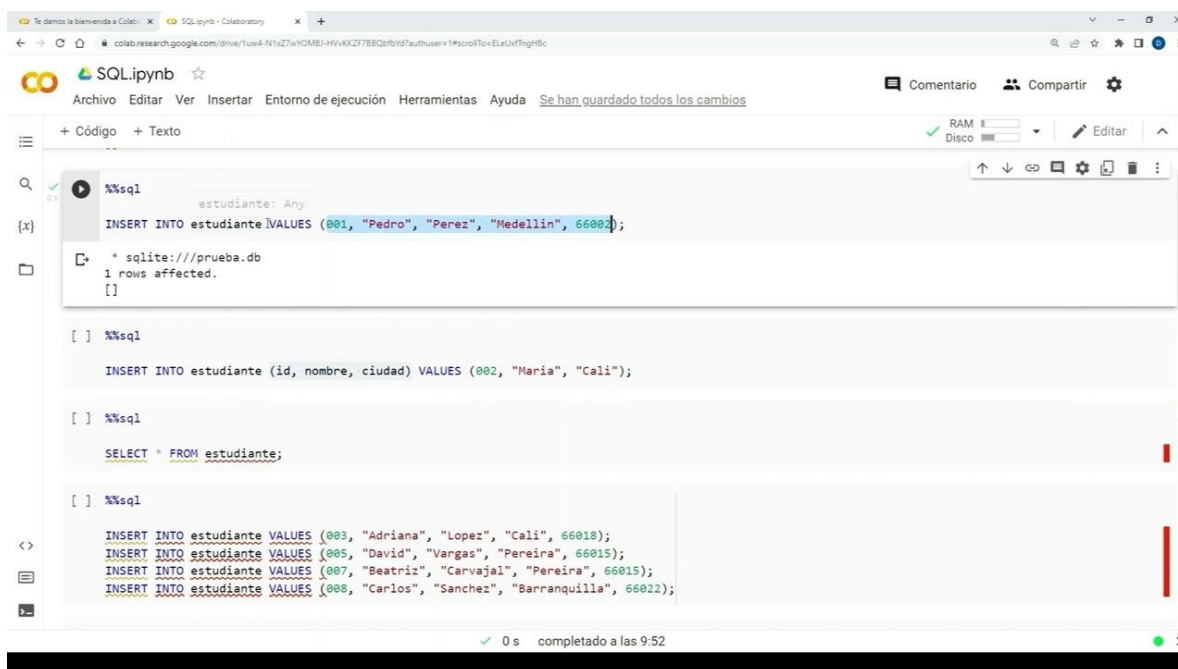
Los sistemas de gestión de bases de datos ofrecen múltiples ventajas, y *MySQL* no es la excepción. Entre las nueve principales ventajas de utilizar *MySQL* se encuentran las siguientes: es de uso libre y gratuito, lo que lo hace accesible para una amplia variedad de usuarios y organizaciones. Además, es un *software* con licencia *GPL (General Public License)*, que define claramente lo que los usuarios pueden o no pueden hacer con el *software*, garantizando así su libre distribución y modificación

bajo los mismos términos. *MySQL* presenta un bajo costo en cuanto a los requerimientos necesarios para la elaboración y ejecución del programa, lo que significa que no se necesita disponer de *hardware* o *software* de alto rendimiento para su funcionamiento. Se caracteriza por su velocidad al realizar operaciones y su buen rendimiento, lo que lo convierte en una opción eficiente para el manejo de grandes volúmenes de datos. La facilidad de instalación y configuración es otro punto a favor, permitiendo a los usuarios implementar el sistema de manera rápida y sencilla. Ofrece soporte en casi el 100% de los sistemas operativos actuales, lo que asegura su compatibilidad y flexibilidad en diversos entornos de trabajo. *MySQL* tiene una baja probabilidad de corrupción de datos, garantizando la integridad y fiabilidad de la información almacenada. Finalmente, proporciona un entorno seguro con capacidades de encriptación, lo que protege los datos contra accesos no autorizados y posibles vulnerabilidades. Estas características hacen de *MySQL* una opción robusta y confiable para la gestión de bases de datos.

*SQL server* pertenece a la casa de Microsoft, con sus herramientas integradas que dan mucho poder para el almacenamiento y proceso de datos.

En el siguiente video se presentan los comandos de *SQL* para la creación de una base de datos y para la realización de consultas en esta:

### Video 8. *SQL* - Comandos *SQL*



```

%%sql
estudiante: Any
INSERT INTO estudiante VALUES (001, "Pedro", "Perez", "Medellin", 66002);

+ sqlite:///prueba.db
1 rows affected.
[]

[ ] %%sql

INSERT INTO estudiante (id, nombre, ciudad) VALUES (002, "Maria", "Cali");

[ ] %%sql

SELECT * FROM estudiante;

[ ] %%sql

INSERT INTO estudiante VALUES (003, "Adriana", "Lopez", "Cali", 66018);
INSERT INTO estudiante VALUES (005, "David", "Vargas", "Pereira", 66015);
INSERT INTO estudiante VALUES (007, "Beatriz", "Carvajal", "Pereira", 66015);
INSERT INTO estudiante VALUES (008, "Carlos", "Sanchez", "Barranquilla", 66022);
  
```

[Enlace de reproducción del video](#)

### Síntesis del video: *SQL* – Comandos *SQL*

Hola, bienvenidos a este video donde podremos ver en la práctica los comandos de *SQL* para la creación de bases de datos y la realización de consultas a la información en la base de datos. Para esta actividad, utilizaremos la herramienta *Google Colab*, una plataforma que nos permite ejecutar comandos de Python en un navegador sin necesidad de instalar *software* adicional ni descargar otros programas. Este es el cuaderno que vamos a utilizar: *Google Colab* trabaja con cuadernos que contienen celdas donde podemos tener tanto texto como código a ejecutar.

En la primera celda de código en Python, instalamos la librería necesaria para ejecutar los comandos *SQL*. Luego, importamos esta librería y creamos el motor de base de datos *SQL*, creando una base de datos llamada "prueba\_db". A continuación, cargamos la extensión de *SQL* y realizamos la conexión a la base de datos creada. Una vez ejecutados estos comandos, la base de datos ya está visible en la carpeta correspondiente.

El primer comando *SQL* que vamos a ejecutar es el de crear tabla. La sintaxis de este comando es la siguiente: "*CREATE TABLE* nombre\_de\_la\_tabla (campo1 tipo\_de\_dato, campo2 tipo\_de\_dato, ...);". En nuestro ejemplo, crearemos una tabla llamada "estudiantes" con cinco atributos: un identificador de tipo entero, un nombre de tipo cadena de caracteres con longitud máxima de 255, apellido, ciudad y código postal (también de tipo entero). Los campos están separados por comas y es recomendable terminar cada instrucción con un punto y coma. Una vez ejecutado el comando, la tabla se crea correctamente.

El siguiente comando es el "*INSERT*", que nos permite agregar información a la tabla creada. Usamos "*INSERT INTO* nombre\_de\_la\_tabla *VALUES* (valor1, valor2, ...);" para enviar los valores correspondientes a los atributos de la tabla. En nuestro caso, enviamos cinco valores para los cinco campos de la tabla "estudiantes". Al ejecutar el comando, se agrega una nueva fila a la tabla. Si solo queremos enviar algunos valores, especificamos a qué campos corresponden dichos valores, por ejemplo: "*INSERT INTO* nombre\_de\_la\_tabla (campo1, campo2) *VALUES* (valor1, valor2);". De esta forma, podemos agregar información parcial a la tabla.

Para consultar la información en la base de datos, usamos el comando *"SELECT"*. Con *"SELECT \* FROM nombre\_de\_la\_tabla;"*, solicitamos todos los campos de la tabla especificada. Si queremos aplicar condiciones a nuestra consulta, usamos la cláusula *"WHERE"*. Por ejemplo, *"SELECT \* FROM estudiantes WHERE ciudad = 'Cali'"* nos trae solo los registros donde la ciudad sea Cali. Además, podemos utilizar comandos de agregación como *"COUNT"* para contar los registros, por ejemplo: *"SELECT COUNT (nombre) FROM estudiantes;"*.

Finalmente, podemos ordenar los resultados de nuestras consultas con el comando *"ORDER BY"*. Usamos *"SELECT \* FROM nombre\_de\_la\_tabla ORDER BY campo;"* para ordenar por un campo específico. Por defecto, el orden es ascendente, pero podemos especificar que sea descendente usando *"DESC"*, por ejemplo: *"SELECT \* FROM estudiantes ORDER BY nombre DESC;"*. Así, los registros se ordenan de manera descendente por el campo nombre.

Espero que este video haya sido útil para entender cómo ejecutar comandos *SQL* en *Google Colab* para la creación y consulta de bases de datos. ¡Gracias por su atención!

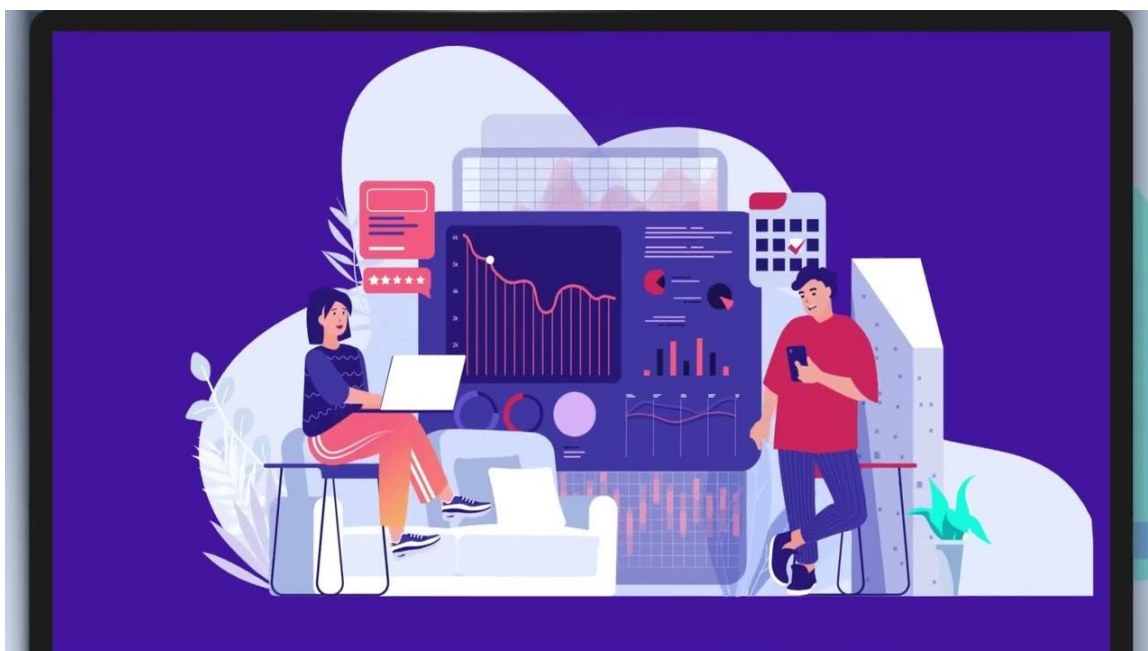
### 5.10. No SQL

Son un tipo de bases de datos cuyo almacenamiento de datos no se realiza en tablas propiamente dicha. Presentan una estructura particular tipo *SON* o *BSON*, que consiste en un arreglo de datos por jerarquías, tiene la ventaja sobre las bases de datos relacionales que su desempeño de búsquedas y cargue de datos son más veloces, además, se podría definir que son más flexibles es sus datos y cambios de estructuras

que se puedan presentar. Para el manejo de *Big data* y grandes cantidades de datos son muy empleadas.

**En siguiente video se explican las ventajas de las bases de Video *No SQL* - Bases de datos relacionales y no relacionales:**

**Video 9. *No SQL* - Bases de datos relacionales y no relacionales**



Video. SENA 2024.

[Enlace de reproducción del video](#)

Síntesis del video: *No SQL* - Bases de datos relacionales y no relacionales

Las bases de datos relacionales y no relacionales son herramientas excelentes para la manipulación de información. Las bases de datos relacionales, como las más conocidas, organizan la información en tablas y permiten consultas y manipulaciones complejas usando el lenguaje *SQL*. Por otro lado, las bases de datos no relacionales,

conocidas como *No SQL (Not Only SQL)*, son ideales para almacenar grandes cantidades de información que no se ajustan al modelo tabular. *No SQL* permite combinar entornos *SQL* y *No SQL* según sea necesario, siendo *SQL* el lenguaje de consulta y manipulación de datos sin necesidad de que estén almacenados en tablas.

Las bases de datos *No SQL* organizan la información como documentos, lo que significa que no es necesario que los datos estén debidamente estructurados para poder manipularlos. Además, estas bases de datos contemplan el uso de tecnologías asociadas a las Industrias 4.0 o *Big Data*, permitiendo manejar grandes volúmenes de datos de manera eficiente y flexible.

### 5.11. *JSON, BSON y XML*

Uno de los desafíos técnicos es la integración entre sistemas de información y la manera en cómo enviar y recibir datos de otras aplicaciones.

Para ello se desarrollaron diversas maneras, entre las más comunes son:

#### ***JSON***

Este estándar se comunica entre sistemas con estructuras y objetos de información básica en código basado en navegador web.

Ejemplo de estructura *JSON*:

De esta manera, se almacenan los datos en Bases de datos *NoSQL* y se intercambian datos entre sistemas de información.



## ***BSON***

Este estándar se comunica entre sistemas con estructuras y objetos de información básica en código basado en navegador web. Ejemplo de estructura *JSON*.

## ***XML***

Son archivos propiamente dicho, está basado en el lenguaje *XML*, usado para guardar e intercambiar datos estructurados a través de sistemas web.

Estos lenguajes y sus archivos derivados mejoran el uso de datos y la forma en que se estructura en Internet, mientras que el *HTML* se encarga del aspecto visual y estilo de la información.

### **5.12. *DDL, DML y DC***

Las bases de datos emplean diferentes tipos de lenguajes y también de allí se derivan archivos necesarios para que los motores de bases de datos funcionen y conserven la información física y lógicamente. Estas son:

#### ***Data Definition Language (DDL)***

Lenguaje de definición de datos: es el encargado de definir estructuras de datos proporcionado por los sistemas gestores de bases de datos permitiendo a los programadores de las bases de datos enfocar y definir estructuras, las más importantes son: *CREATE* (crear tablas o bases de datos), *ALTER* (cambiar o redefinir tablas o campos), *DROP* (Limpiar tablas, eliminar objetos, índices, etc.).

#### ***Data Manipulation Language (DML)***

Lenguaje de Manipulación de Datos: permite a los motores de datos instrucciones de *SQL*, otorga a los usuarios introducir datos para luego ejecutar tareas

de consultas o modificación de los datos en las tablas especificadas. Los elementos que se utilizan para manipular los datos son:

- *SELECT*. Sentencia para realizar consultas sobre los datos.
- *INSERT*. Permite insertar valores en una tabla.
- *UPDATE*. Modificar los valores de uno o varios registros.
- *DELETE*. Eliminar registros de una tabla.

### ***Data Control Language] (DCL)***

Lenguaje de Control de Datos: permiten administrar base de datos, controlar el acceso a los objetos, es decir, podemos otorgar o denegar permisos a uno o más roles para realizar determinadas tareas. Los comandos para controlar los permisos son los siguientes:

- *GRANT*. Permite asignar permisos.
- *REVOKE*. Elimina los permisos asignados.

## **6. Estructuras y componentes de analítica de datos**

A continuación, este apartado se emplea para dar una breve introducción a elementos muy importantes a hora de aplicar inteligencia de negocios. Se describe a modo conceptual, algunos elementos de las bodegas de datos y sus topologías.

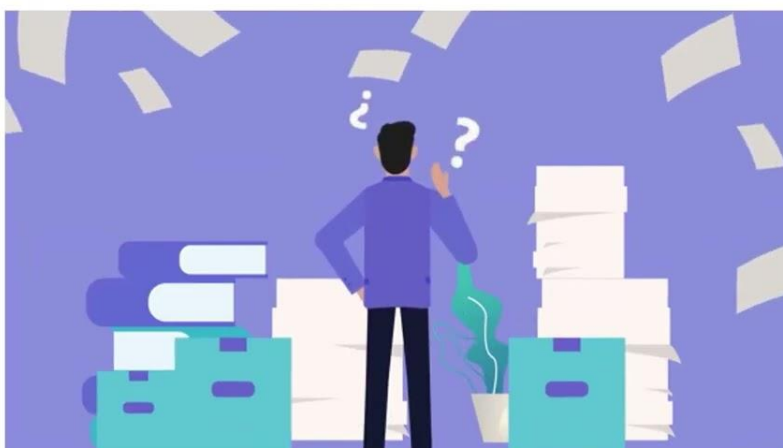
La estructura completa de la analítica tiene dos extremos donde fluyen y se transforman los datos, desde los orígenes, pasando por un proceso de carga al sistema analítico, limpieza de datos, optimización de datos para ser guardados de forma consolidada y completa en un arreglo de bases de datos denominado Bodega de datos,

pero acá no acaba el flujo, el final es la explotación de datos traducidos en reportes, cubos de datos u otros usos como minería de datos.

Con el siguiente video, se amplía información sobre analítica de datos, con lo cual puede despejar algunas dudas:

### Video 10. Analítica de datos

## ANALÍTICA DE DATOS



**Para  
qué?**



[Enlace de reproducción del video](#)

#### Síntesis del video: Analítica de datos

Hola a todos, el día de hoy vamos a hablar de analítica de datos. Inicialmente, vamos a conocer la definición. ¿Qué entendemos por análisis de datos? Lo definimos como el proceso de recoger información y luego analizarla para confirmar varias hipótesis. El análisis de datos también significa contar historias con datos, transmitir de forma clara y concisa el estado del mundo a quienes nos rodean. Es el uso de la

información que nos rodea para tomar decisiones, al igual que cuando te levantas cada mañana, ves las noticias y el informe meteorológico que indica la temperatura del día y si va a llover, lo que puede influir en lo que vas a llevar puesto y las actividades que puedes realizar.

El análisis de datos no es un concepto abstracto, es algo que hacemos de forma natural, pero tiene un nombre técnico y ahora se paga a la gente para que lo realice como una experiencia mucho mayor o grandiosa. Usamos la forma en que lo digo, que hay un problema y que necesitamos usar los hechos para probar una hipótesis. Ahí es donde entra en juego el análisis de datos. El proceso comienza con la definición del problema y luego tienes que crear una propia hipótesis para probar. Para esto hay que recoger datos, limpiar datos, analizar las bases y luego presentarlos a los principales interesados.

El análisis de datos es realmente cualquier conjunto de datos que puedas usar para revisar información, cualquier cosa que ayude a entender lo que está pasando. Para qué siempre estamos analizando datos en la vida cotidiana, para predecir dónde ha estado alguien, dónde está ahora mismo y hacia dónde se dirige. Estos datos me ayudan a ver más allá y casi a predecir el futuro de cualquier compañía con la que trabajo. El análisis de datos es recolectar, limpiar, analizar, presentar y, en última instancia, compartir los datos y los análisis para poder ayudar a comunicar exactamente lo que está pasando en la empresa. Lo que está pasando con los datos puede ayudar a tomar mejores decisiones para una gestión organizacional.

Podemos concluir que el análisis de datos es un proceso, o mejor aún, un fenómeno de tomar información recopilada de una población relevante, tal vez clientes, audiencia social, proveedores, conjuntos y usar estos datos para tomar decisiones sobre productos o servicios que queremos ofrecer o mejorar. En el entorno digital en el que nos encontramos actualmente, el buen uso del análisis de datos permitirá a las empresas lograr un camino donde podrán ser más atractivas para los clientes, tomar decisiones en sus procesos, brindando beneficios claros, reduciendo costos y optimizando el desempeño de todos sus procesos. Muchas gracias.

### **6.1. Bodega 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 entonces una estructura diseñada y desarrollada para almacenar y procesar datos de múltiples fuentes y centralizarlas para la elaboración de reportes y datos analíticos.

Las bodegas de datos son el corazón de la inteligencia de negocios, allí se almacena de manera incremental toda la información producida por la organización, cuando se establecen desarrollos completos de analítica de datos, se hace necesario

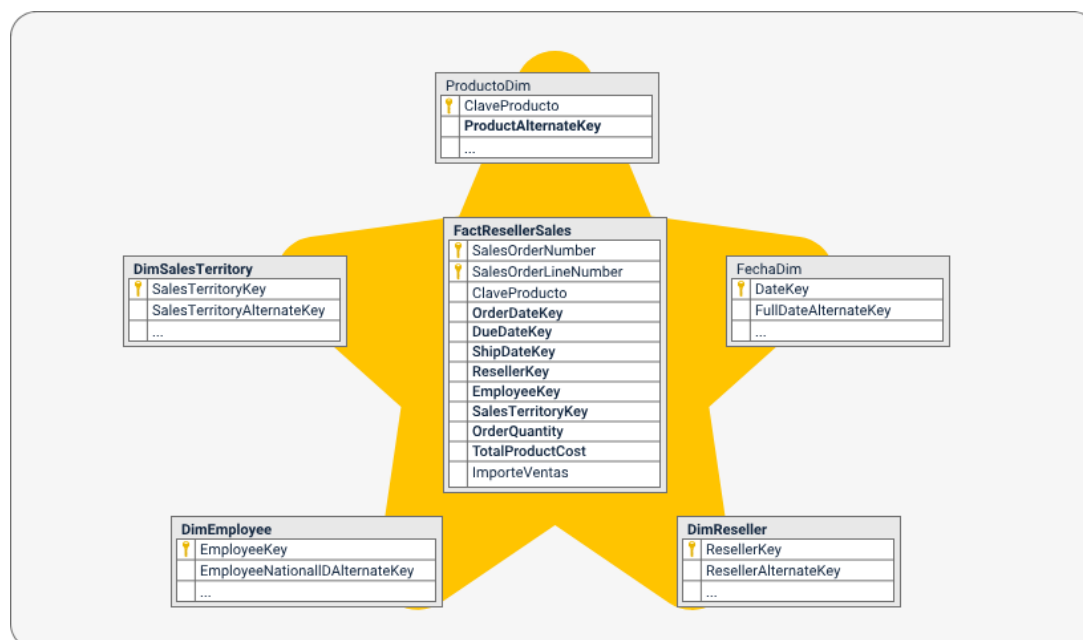
contar con infraestructura y gestión de almacenamiento igual de robustas que den respuesta a las exigencias de la cantidad y variedad de datos que se presentan. Para grandes empresas, es posible que se cuente con especialistas para cada proceso del flujo de datos, así mismo el almacenamiento de las bodegas de datos requiere 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 de tal manera que se pueda visualizar de una mejor manera, las tablas que componen la bodega de datos se pueden presentar en diversos diseños, como tipo estrella, copo de nieve, constelaciones, etc.

## 6.2. Tipos estrella

Es aquella que cuyas dimensiones se relacionan directamente con la tabla de hechos, se representa de la siguiente manera (explorar figura):

**Figura 6. DWH Tipo Estrella**

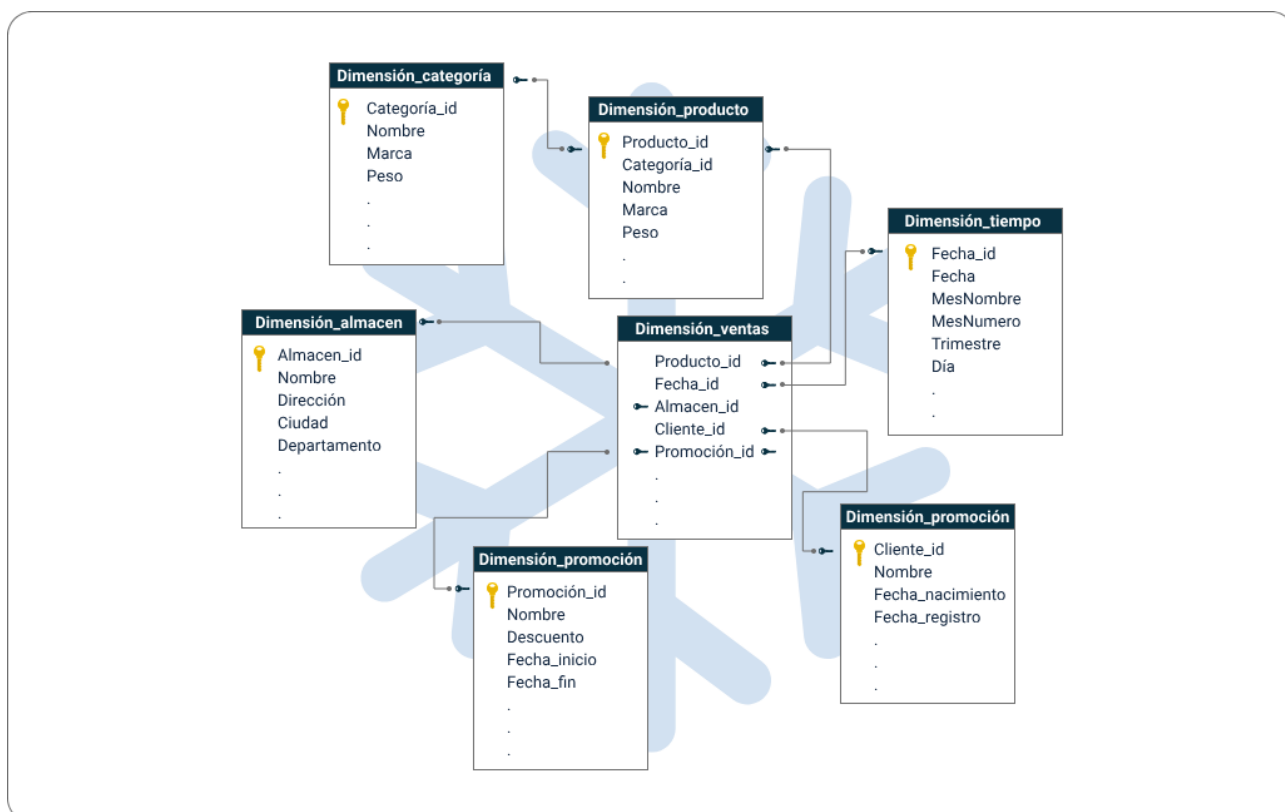


Vale la pena señalar que muchas tablas giran en torno a la tabla principal, conocida como la tabla de hechos, la cual generalmente contiene las claves primarias de las dimensiones asociadas, así como las cifras o medidas obtenidas previamente en el proceso *ETL*.

### 6.3. Copo de nieve

La topología copo de nieve o *Snowflake* se presenta frecuentemente, cuando las dimensiones cuentan con otras familias o categorías que obligan a incluir subdimensiones para completar la información. Ejemplo de esto es cuando existe una dimensión Productos, es muy frecuente que exista la dimensión Categorías. Esta última se conecta a la tabla de hechos a través de Productos que está en medio. Se representa de la siguiente manera (explorar figura):

**Figura 7. DWH Snowflake**



En este caso, al igual que con la topología de estrella, los hechos se relacionan con las dimensiones a partir de las claves de las tablas que lo circundan, sin embargo, en este caso se pueden presentar más niveles donde una dimensión pueda contener una clave foránea de otra tabla. En ese sentido, como se ve en el ejemplo, la dimensión\_categoria se asocia a los hechos a través de una tabla en medio llamada dimension\_producto.

#### **6.4. Constelación**

Similar a la topología copo de nieve, sin embargo, se presenta cuando hay más de dos niveles de relación. Es decir, hay más de una tabla en medio entre una dimensión y la tabla de hechos. Esta arquitectura no es muy común y es poco eficiente, pues ya los diseños dimensionales se van desdibujando un poco y comienzan más a parecerse a estructuras transaccionales que multidimensionales.

Al sumar muchos datos en tablas de hechos y dimensiones, más el procesamiento que implica la lectura y escritura de datos con muchas relaciones de tablas y todo de manera masiva, no hace muy veloz estas arquitecturas, ocasionalmente algunas soluciones traen las tablas del sistema transaccional a las bodegas de datos sin procesos *ETL*, implicando mucha carga para los sistemas de reportes ocasionando su colapso en algunas herramientas de visualización de datos.

### **7. Herramientas para el análisis de datos**

Para el proceso de analizar los datos se utiliza una serie de herramientas que se comparten en la siguiente tabla con su función principal:



**Tabla 3.** Variables y escalas

| Herramientas para el análisis de los datos |   |
|--|---|
| Herramientas                               | Función   |
| <b>Microsoft Power BI</b>                  | Es una herramienta de análisis segura que proporciona una vista interactiva de la información, dando acceso a más de 60 fuentes y compatible con otras aplicaciones.  |
| <b>Programación en R</b>                   | Esta herramienta de análisis permite la estadística y su modelación, se adapta a varias plataformas por medio de más de 11.000 paquetes que se instalan acorde a cada necesidad de forma automática.  |
| <b>SAS</b>                                 | Esta herramienta de análisis actúa como lenguaje en la programación, permitiendo que la información sea procesada de forma separada, siendo útil en la gestión de perfilamiento de los clientes, predicción de compras y demás.                             |
| <b>Python</b>                              | Es una herramienta diseñada para trabajar sobre objetos y procesar datos de forma funcional y estructurada.   |
| <b>Excel</b>                               | Esta herramienta utilizada por la gran mayoría de empresas es básica; pero muy útil para analizar los datos de los clientes y se puede ajustar gracias a que cuenta con fórmulas internas que permiten generar frecuencias, filtros, combinaciones y demás. |
| <b>Rapid Miner</b>                         | Es una herramienta para realizar análisis predictivos.  |
| <b>Apache Spark</b>                        | Es una herramienta que procesa los datos de forma rápida, con algoritmos que le permiten clasificar la información.   |

| Herramientas para el análisis de los datos |   |
|--|---|
| Herramientas                               | Función   |
| <i>Qlik View</i>                           | Esta herramienta procesa la información comprimiéndola, ahorrando espacio en el disco duro y la asocia relacionando la información por colores según se requiera. |

Nota: las herramientas vistas en la tabla anterior son buenas opciones para la gestión del análisis de los datos; pero su uso depende de las necesidades de cada entidad.

### 7.1. Entornos de desarrollo – IDE

En este apartado, se mencionan los Entornos de Desarrollo Integrado (*IDE*), los cuales ahorran mucho tiempo y esfuerzo a quienes programan en la preparación de las plataformas e instalación de complementos extras que requieren, de esa manera, solo se dedican a programar y dejar todos los recursos necesarios a que se incluyan en estos entornos. Para la gestión de datos se emplean muchos lenguajes de programación, y lenguajes de consulta, lo que implica prácticamente construir códigos enteros para los procesos de datos.

### 1. Google *collaborate*

Entorno completo en línea de Google. Es versátil, fácil y es muy usado en contextos académicos y aprendizaje de lenguajes de programación.

Sitio: <https://colab.research.google.com/>



### 2. *Jupyter*

Puede ejecutarse en Google *collaborate*, pero tiene su propio entorno llamado *JupyterLab*.

Sitio: <https://jupyter.org/>



### 3. *PyCharm*

Editor de código muy potente multilenguaje.

Sitio: <https://www.jetbrains.com/es-es/pycharm/>



#### 4. Anaconda

Muy empleada para desarrollar lenguaje R, *Python*. Maneja grandes volúmenes de datos y diferentes análisis.

Sitio: <https://www.anaconda.com/products/distribution> o <https://youtu.be/xrIISRh0MZs> (instructivo para instalar anaconda).



#### 5. Orange

Un *IDE* con recursos de *ML*, visualización de datos. Como la mayoría de código abierto.

Sitio: <https://orangedatamining.com/download/>



#### 7.2. Python

Actualmente, es el lenguaje de programación que lidera los desarrollos basados en gestión de datos. En relación con otros lenguajes presenta curva de aprendizaje rápido, cuenta con múltiples librerías que expanden su capacidad y muchos sistemas de manejo de datos e inteligencia artificial lo emplean para codificar sus funcionalidades.

Para desarrollar aplicaciones y realizar minería de datos basado en inteligencia artificial puede emplear alguno de los *IDE* anteriormente conocido u otros *Code Skulpor*

[Enlace web](#)

### 7.3. Librerías

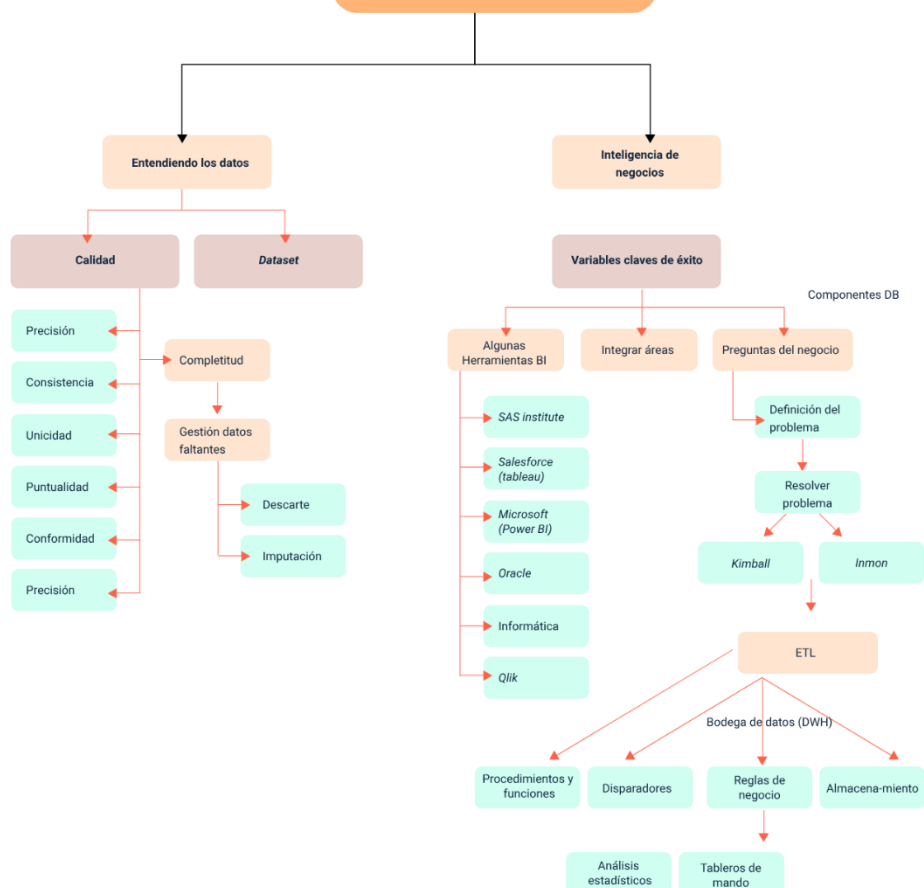
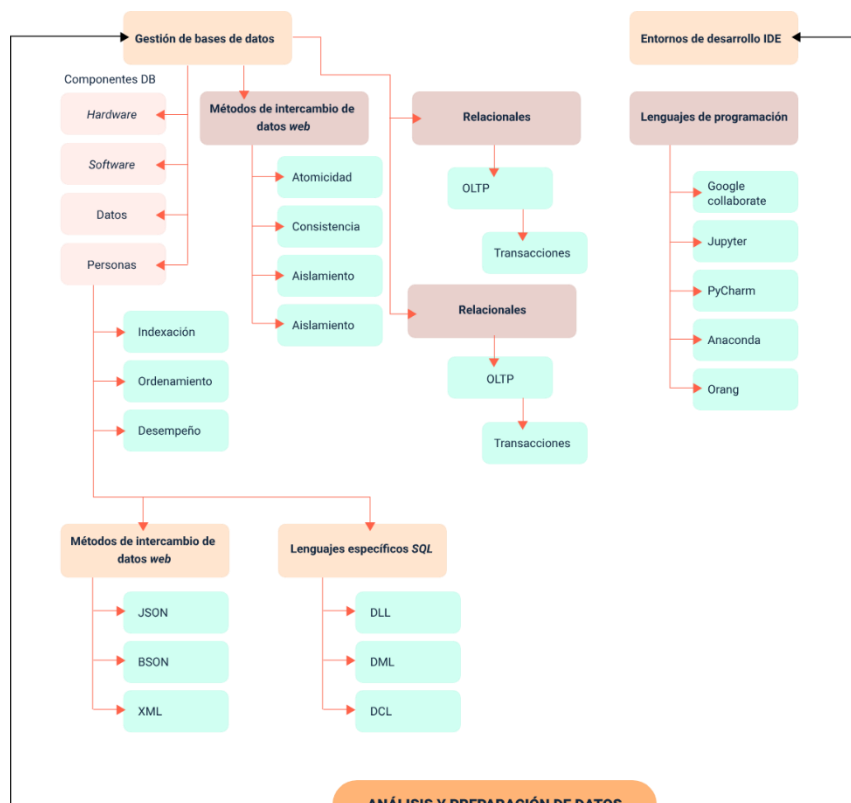
Las librerías en los lenguajes de programación son un conjunto de archivos que contienen códigos de programas o funcionalidades que hacen tareas concretas y repetitivas para facilitar la programación. Las librerías se categorizan por el tipo de funcionalidades que presentan, algunas son de tipo estadístico, otras para la visualización de datos, otras para operaciones matemáticas especializadas, etc.

**En Python, las más comunes y útiles son:**

- **Pandas:** para la ciencia de datos es una librería muy usada, pues facilita la manipulación y consulta de datos.
- **Numpy:** permite generar una estructura de datos universal, lo que se traduce un mejor análisis de datos, y emplea algoritmos muy poderosos para el intercambio de estos datos. Tiene buen desempeño en datos masivos (dependiendo también del *hardware*).
- **Matplotlib:** con esta librería se generan gráficos de calidad para publicar en línea o en archivos como PDF y sin emplear muchas líneas de código. Se pueden generar gráficos de barras, histogramas, series temporales, espectros de potencia, entre muchas más visualizaciones.

## Síntesis

En el siguiente diagrama se expone la síntesis de la complejidad de la gestión y preparación de datos, destacando componentes como *hardware*, *software*, datos y personas. Se abordan métodos de intercambio de datos web, incluyendo *JSON*, *BSON* y *XML*, y principios de transacciones como atomicidad y consistencia, asimismo se detallan lenguajes *SQL* específicos (*DDL*, *DML*, *DCL*) y entornos de desarrollo IDE como *Google Collaborate*, *Jupyter* y *PyCharm*, resaltando las interconexiones necesarias para un análisis y preparación de datos efectivos:



## Material complementario

| Tema  | Referencia APA del material   | Tipo                    | Enlace  |
|---|---|-------------------------|---|
| 1.2 Técnicas de almacenamiento de datos y consultas | Ecosistema de Recursos Educativos Digitales SENA. (2021). Aplicando el MER con herramienta Día    Cardinalidad modelo entidad relación. SENA. | Video                   | <a href="https://www.youtube.com/watch?v=KcORnp2A3yg">https://www.youtube.com/watch?v=KcORnp2A3yg</a>   |
| 2.2 Detección de errores y datos faltantes          | Codificandobits. (s.f). Mapa paso a paso manejo datos faltantes. Blog.  | Manual de procedimiento | <a href="https://www.codificandobits.com/descargas/dl_202100618_mapa_pasos_manejo_datos_faltantes.pdf">https://www.codificandobits.com/descargas/dl_202100618_mapa_pasos_manejo_datos_faltantes.pdf</a> |
| 2.3 Identificación de variables importantes         | Sotaquirá, M. (2021). Guía completa para el manejo de datos faltantes. Blog.  | Artículo                | <a href="https://www.codificandobits.com/blog/manejo-datos-faltantes/">https://www.codificandobits.com/blog/manejo-datos-faltantes/</a>   |
| 3.7 Procedimientos almacenados y funciones          | Calbimonte, D. (2019). Funciones frente a los procedimientos almacenados en SQL Server.   | Tutorial                | <a href="https://www.sqlshack.com/es/funciones-frente-a-los-procedimientos-almacenados-en-sql-server/">https://www.sqlshack.com/es/funciones-frente-a-los-procedimientos-almacenados-en-sql-server/</a> |
| 4. Análisis exploratorio de datos                   | Codificandobits. (s.f). Guía paso a paso análisis exploratorio. Blog.   | Mapa conceptual         | <a href="https://www.codificandobits.com/descargas/dl_202100611_mapa_pasos_analisis_exploratorio.pdf">https://www.codificandobits.com/descargas/dl_202100611_mapa_pasos_analisis_exploratorio.pdf</a>   |
| 5.8 Algebra relacional                              | Cidecam. (2021). Algebra Relacional. Página web.  | Artículo                | <a href="http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro14/33_algebra_relacional.html">http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro14/33_algebra_relacional.html</a>                         |
| 7.3 Librerías                                       | Ecosistema de Recursos Educativos Digitales SENA. (2024). Manejo de datos con Pandas. SENA.   | Vídeo                   | <a href="https://www.youtube.com/watch?v=qanhqbz7pME">https://www.youtube.com/watch?v=qanhqbz7pME</a>   |



## Glosario

### I

Información Descriptiva: se refiere a las historias que los datos cuentan, se refiere a un histórico de eventos y resultados.

Información Predictiva: se refiere a los hechos probables que sucederán, esto se realiza basado en datos descriptivos, es decir, datos históricos y procesos matemáticos y/o modelación de *Machine Learning*.

### J

*Joins*: proceso de selección de datos de varias tablas y unirlos en una vista o nueva tabla. Una instrucción de “*SQL JOIN*” en un comando *Select* que combina columnas entre una o más tablas en una base de datos.

### L

Lenguaje *DAX*: lenguaje específico para gestión de datos creado por Microsoft (*Data Analysis Expressions*). Se emplea en colecciones de datos en aplicaciones como Excel, *Analysis Services* y *Power BI*.

Lenguaje R: lenguaje de programación para la gestión de datos. Es un lenguaje interpretado que ejecuta las instrucciones directamente sin previa compilación.

### N

Normalización: la estructura organizada en datos relacionales que cumplen unas reglas de normalización que garantizan la integridad, calidad y optimización en la base de datos.

## P

Procesadores: *CPU* (Unidad central de proceso), es el componente del computador y otros dispositivos programables, que interpreta las instrucciones contenidas en los programas y procesa los datos.

Procesamiento por lotes: al tener muchas cargas de datos y procesamientos, es necesario reunir recursos para que se ejecuten de manera independiente optimizando recursos, de esta manera las tareas se completan periódicamente de manera repetitiva.

## R

*RAM*: es la memoria temporal o de corto plazo de las computadoras, es la memoria principal de trabajo, los programas y datos se cargan allí para que trabajen más rápidamente.

## S

*Script*: se refiere a fragmentos de código de programación que pueden ejecutar una o varias funciones.

Sistema operativo: es el *software* principal de las computadoras, se emplea como plataforma para gestionar las aplicaciones, recursos del *hardware* y entornos gráficos y funcionales.

## T

Tabular: en estadística, son la recopilación y procesamiento de la información capturada de los instrumentos disponibles al momento de realizar encuestas, toma de datos y otras.

*Ti:* (IT) Abreviatura de Tecnología de la información.

## Referencias bibliográficas

Banco de la República. (2022). Sistema de información económica de la Gerencia Técnica.

[https://totoro.banrep.gov.co/analytics/saw.dll?Portal&PortalPath=%2Fshared%2FDashboards\\_T%2FD\\_Estad%C3%ADsticas%2FEstad%C3%ADsticas&NQUser=publico&NQPassword=publico123&lang=es&page=Precios%20e%20inflaci%C3%B3n](https://totoro.banrep.gov.co/analytics/saw.dll?Portal&PortalPath=%2Fshared%2FDashboards_T%2FD_Estad%C3%ADsticas%2FEstad%C3%ADsticas&NQUser=publico&NQPassword=publico123&lang=es&page=Precios%20e%20inflaci%C3%B3n)

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

Curto Díaz, J. (2016). Organizaciones orientadas al dato: transformando las organizaciones hacia una cultura analítica. Editorial UOC. <https://elibro-net.bdigital.sena.edu.co/es/lc/senavirtual/titulos/58609>

Fernández, J. (2021). Escalas de medición de las variables: nominal, ordinal, intervalo y razón. <https://youtu.be/XNulqSfCskQ>

Gawande, S. (2020). iCEDQ Torana INC. Obtenido de 6 Dimensions of Data Quality, Examples, and Measurement: <https://icedq.com/6-data-quality-dimensions>

Ommi, A. K. (18 de 02 de 2018). Introduction to Data and Information. MyCloudWiki: <https://www.mycloudwiki.com/san/data-and-information-basics/>

Pang, A., Markovski, M., & Ristik, M. (22 de septiembre de 2022). Top 10 Analytics and BI Software Vendors, Market Size and Market Forecast 2023-2028. Apps Run the World: <https://www.appsruntheworld.com/top-10-analytics-and-bi-software-vendors-and-market-forecast/>

Pulido Romero, E., Escobar Dominguez, O., & Núñez Pérez, J. (2019). Bases de datos.

México DF: Grupo Editorial Patria. Obtenido de <https://elibro-net.bdigital.sena.edu.co/es/lc/senavirtual/titulos/121283>

Velthuis, M. P. (2019). Calidad de datos. Bogotá: Ediciones de la U. <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=9094>

## 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 de la Industria, la Empresa y los Servicios CIES |
| Giovanna Andrea Escobar Ospina | Diseñadora Instruccional          | Regional Norte de Santander - Centro de la Industria, la Empresa y los Servicios CIES |
| Silvia Milena Sequeda Cárdenas | Asesora Metodológica y Pedagógica | Regional Distrito Capital - Centro de Diseño y Metrología                             |
| Julia Isabel Roberto           | Corrección de Estilo              | Regional Distrito Capital - Centro de Diseño y Metrología                             |
| Sandra Paola Morales Páez      | Evaluadora Instruccional          | Regional Santander - Centro Agroturístico   |
| Jaime Hernán Tejada            | Experto Temático                  | Regional Norte de Santander - Centro de la Industria, la Empresa y los Servicios CIES |

## DISEÑO Y DESARROLLO DE RECURSOS EDUCATIVOS DIGITALES

|                               |                                    |   |
|-------------------------------|------------------------------------|---|
| Yuly Andrea Rey Quiñonez      | Diseñadora de Contenidos Digitales | Regional Santander - Centro Agroturístico |
| Lizeth Karina Manchego Suarez | Desarrolladora Full-Stack          | Regional Santander - Centro Agroturístico |
| María Alejandra Vera Briceño  | Animadora y Productora Multimedia  | Regional Santander - Centro Agroturístico |

## VALIDACIÓN RECURSO EDUCATIVO DIGITAL

|                                 |  |   |
|---------------------------------|--|---|
| Yineth Ibette González Quintero | Validadora de Recursos Educativos Digitales        | Regional Santander - Centro Agroturístico |
| Laura Paola Gelvez Manosalva    | Validadora de Recursos Educativos Digitales        | Regional Santander - Centro Agroturístico |
| Erika Fernanda Mejía Pinzón     | Evaluadora Para Contenidos Inclusivos y Accesibles | Regional Santander - Centro Agroturístico |