**FORMATO PARA EL DESARROLLO DE COMPONENTE FORMATIVO**

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

| COMPETENCIA | 220501114- Sistematizar datos masivos de acuerdo con métodos de analítica y herramientas tecnológicas. | RESULTADOS DE APRENDIZAJE | 220501114-01 - Preparar datos de según las fuentes de información. |
| --- | --- | --- | --- |

| NÚMERO DEL COMPONENTE FORMATIVO | CF13 |
| --- | --- |
| NOMBRE DEL COMPONENTE FORMATIVO | 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. |
| PALABRAS CLAVE | Analítica, Datos, Inteligencia de negocio, Origen de datos |

| ÁREA OCUPACIONAL | VENTAS Y SERVICIOS |
| --- | --- |
| IDIOMA | Español |

1. **TABLA DE CONTENIDOS:**

**Introducción**

**1. Modelamiento y gestión de bases de datos**

1.1 Metodologías de diseño y normalización Principios de ACID

1.2 Técnicas de almacenamiento de datos y consultas

1.3 Estándares técnicos

1.4 Ordenamiento de datos, indexación y recuperación

**2. Preparación de datos**

2.1 Entendimiento de la data

2.2 Detección de errores y datos faltantes

2.3 Identificación de variables importantes

2.4 Dataset

**3. La inteligencia de negocios**

3.1 Identificación de las preguntas básicas

3.2 Metodología de integración

3.3 Herramientas de administración

3.4 Técnicas de solución de problemas (modelación de datos)

3.5 Metodologías de análisis (Kimball, Inmon)

3.6 Verificación de valores y escalas

3.7 Procedimientos almacenados y funciones

3.8 Disparadores

**4. Análisis exploratorio de datos**

4.1 Estadística descriptiva y estadística inferencial

4.2 Población y muestra

4.3 Escalas de medida y clasificación de variable

4.4 Técnicas de análisis estadístico

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

5.1 Datos univariables

5.2 Datos bivariados

5.3 Datos multivariables.

5.4 Reglas de negocio

5.5 Tipo de restricciones

5.6 Programación transaccional

5.7 Programación de estructuras no lineales, desnormalización, series y dataframes

5.8 Álgebra relacional

5.9 SQL

5.10 No SQL

5.11 JSON, BSON y XML

5.12 DDL, DML, DCL

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

6.1. Bodega de datos

6.2 Tipos estrella

6.3 Copo de nieve

6.4 Constelación

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

7.1 Entornos de desarrollo - IDE

7.2 Python

7.3 Librerías

1. **INTRODUCCIÓN**

En este componente se abordarán los conceptos y fundamentos de la inteligencia de negocios, con el fin de realizar el modelamiento, análisis y preparación de los datos. Veamos un video que nos contextualiza al respecto:



1. **DESARROLLO DE CONTENIDOS:**

**1. Modelamiento y gestión de bases de datos**

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. **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 estropeara la aplicación y viceversa.

| A mediados de la década de 1970s, 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. | Centro de datos |
| --- | --- |

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 Dominguez, & 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 la cuenta del comprador y exista 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, al fallar 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), que son:



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

En la actualidad, se presentan diferentes modelos para la organización y de datos; en este caso 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. |  |
| --- | --- |

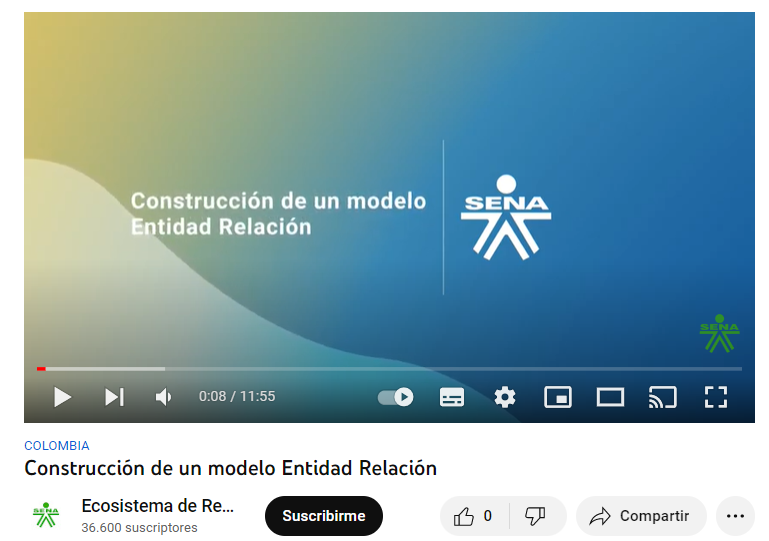
Desde el enfoque de la base de datos, se refiere al uso al que se le va a dar a los datos que se almacenan, en este sentido se pueden mencionar principalmente dos:

* **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 que entidad se definen 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 desarrollo que dan solución a un proceso específico del negocio, como Sistema de facturación, sistema contable, sistemas de inventarios y compras, etc. Se denomina, **OLTP** en español 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 creación de un modelo entidad relación:



* **Bases de datos Dimensionales (OLAP)**: 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:



**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 convenciones para bases de datos son:

* Los nombres de las bases de datos deben contener el nombre de la empresa seguido del nombre de la aplicación.
  + Company\_nombreaplicacion
* Las tablas se deben nombrar en plural, y debe estar asociado a los datos que se almacenarán.
  + facturas, productos, detalle\_facturas
* Por normalización, todas las tablas deben tener una llave primaria la cual debe ser id.
  + facturas -> facturas\_id. productos -> productos\_id . usuario -> 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.

* Los campos de la tabla deben estar en singular y deben describir los datos almacenados.
  + usuarios (nombre, email, clave, fecha\_nacimiento…)
* 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 misma.

Por eso en importante comprender el concepto de **“INDEXACIÓN”,** imagina, tomar un libro, del cual quieres leer un tema específico, ¿cuánto 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 llegarás 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. | Administrador de archivos y concepto de almacenamiento de datos de computación en la nube vector plano aislado |
| --- | --- |

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, la buscaría en toda la tabla, mientas si se usan los punteros 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, los cuales son:

* **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, es usada si tenemos bases de datos que su búsqueda sea por temáticas o que el usuario tenga solo 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 emplea estos comandos:
    - CREATE INDEX nom\_indice ON tabla(nom\_campo);
    - CREATE INDEX idx\_apellidos 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\_nombresCompletos ON usuarios(nombre, 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, muestra 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 muestra, se puede realizar el siguiente ejercicio:

| Con la sentencia  SELECT \* FROM productos    Muestra todos los campos y datos de la tabla productos |  |
| --- | --- |

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    Muestra todos los campos que cumplen con la condición. |  |
| --- | --- |

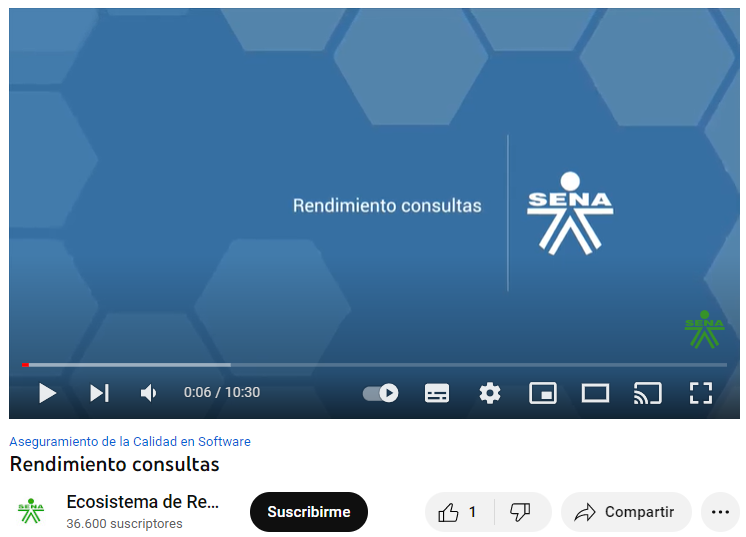
A la consulta anterior, al adicionarle EXPLAIN al inicio, dará un reporte de cómo se realizará la consulta:

| *Ejecución del comando Explain:*  Muestra 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*. |  |
| --- | --- |

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  Muestra 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, te invito para que amplíes tus visualizando el siguiente video:



**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 configurarse los datos a través del proceso comunicativo elemental, (transmisor, emisor, canal, MENSAJE [Información] y contexto), la información y los datos empiezan a surgir en diversas maneras, incluso miles de mensajes de las primeras civilizaciones han quedado guardados por miles de años hasta nuestros días a través de pinturas, esculturas y símbolos. | Talla jeroglífica en un sarcófago egipcio antiguo en un foco selectivo del templo antiguo |
| --- | --- |

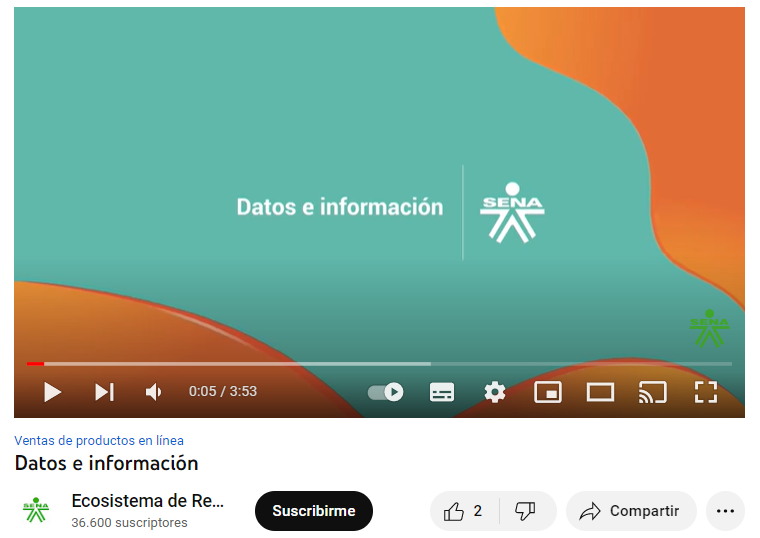
Bajo este contexto, es importante recordar la diferencia entre datos e información. Cuando las civilizaciones descubren mensajes de humanos antiguos, no hay manera de interpretar un mensaje, ante los ojos no entrenados no es más que símbolos con valor artístico, quizás con la cognición antropológica de que los mensajes representan cosas, pero realmente serían solo datos (símbolos) sin información (no interpretados).

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

|  | 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 que genera la información dada por los datos. |
| --- | --- |

Bajo el contexto técnico, los datos en la era moderna se guardan a través de máquinas digitales bajo una herramienta y concepto denominado Bases de datos (BD).

Le invitamos a visualizar el siguiente video donde se profundiza en la diferenciación entre dato e información:



**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. 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, equivalente a elementos microscópicos en medio de los océanos de información es 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 que 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.

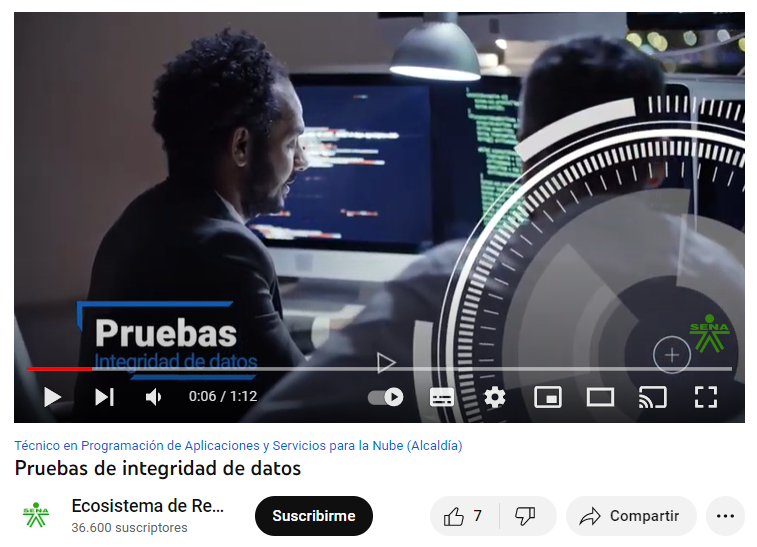
La calidad de los datos toma aún mayor fuerza para aplicar validaciones con regularidad, cuando se habla de ecosistemas de datos, donde convergen diversas infraestructuras de datos donde se interrelacionan datos compartidos.

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 adicionamos otras dimensiones más: Actualizado, Conformidad, Integridad y Precisión, que complementan las dimensiones DQ.

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



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



**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 para garantizar un óptimo resultado.

Cuando los sistemas de información no son capaces de digitalizar eficaz y eficientemente la realidad del negocio, capturándola con el nivel adecuado de detalle necesario, almacenar dichos datos garantizando que no se producen pérdidas sintácticas ni semánticas, procesarlos de acuerdo con las reglas de negocio, o mostrar 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 de datos y la determinación de datos faltantes se debe tener presente que en gran parte del tiempo que se invierte en los procesos de analítica es la de la limpieza, depuración y mejorar 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 estos datos, y en vista de que no existe manera de calcularlos o sacarlos de otras fuentes, los datos que faltan en medio del conjunto de datos se pueden gestionar, principalmente con las dos siguientes opciones:

* **Descarte:** es la más sencilla, se puede tomar de dos maneras:



* **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 aplicarla imputación o reemplazo del dato:



**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:



**2.4 Dataset**

Se denomina al conjunto de datos, lo que traduce en una colección relacionada 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 de 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 exporta e incluye 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 cómo se almacenan y gestionan los datos digitales, estos pueden clasificarse en términos generales como Datos estructurados, semiestructurados (Ommi, 2018).

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



La mayoría de los datos generados por todos los usuarios están clasificados como No estructurados, esto indica que se compone de información difícil de clasificar y ordenar, está dada 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, podrían 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. | Ilustración de concepto informe de datos |
| --- | --- |

En forma simple, la inteligencia de negocios es una serie de técnicas, metodologías y herramientas que se integran para convertir los datos en información, estos 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 aterrizar a los directivos en hacer conciencia de que muchos problemas de la organización se deben a la falta de datos al instante, actualizados y que reflejan una realidad.

| Ilustración del concepto de plan de negocios | Una de las condiciones iniciales más importantes son las preguntas, es decir, qué se necesita saber del negocio, qué decisiones se planean tomar y qué insumos se requieren para la toma de estas decisiones. |
| --- | --- |

Otro concepto errado frecuente, en especial las pymes, es que el indicador más importante y en algunos casos el único, es el dinero que ingresa a la empresa, si bien 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, debería reflejarse 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 iniciarse bajo necesidades reales, preguntas fundamentales y 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 en como BI integra todas o varias áreas del negocio:



**3.3 Herramientas de administración**

Las herramientas disponibles para la implementación de Inteligencia de negocio en las organizaciones, es amplia 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: Que 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:



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 alguna información 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 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 actual 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**

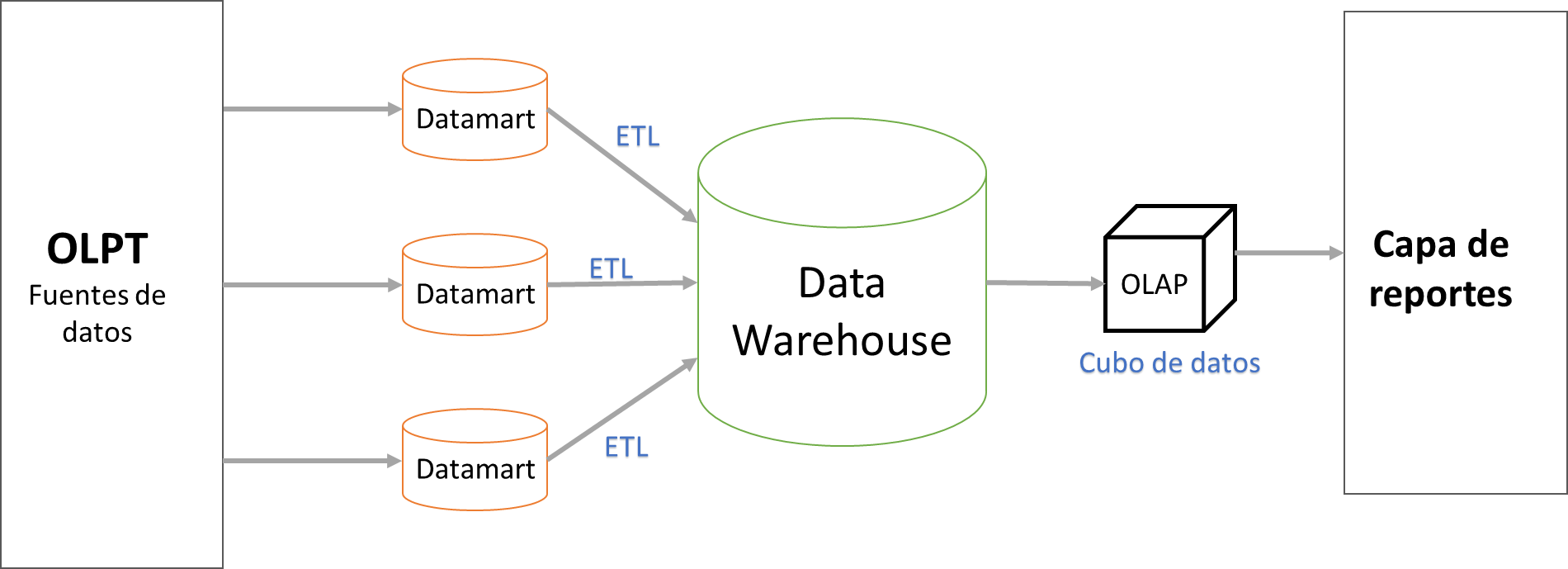
La metodología para el diseño de DWH propuesta por *Ralph Kimball* se centra en copiar las bases de datos transaccionales en un modelo optimizado para consultas de analítica. Para integrar varias áreas o disciplinas del negocio se emplean *data marts*, que son bases de datos que surgen a partir del proceso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

Figura 1.

***Modelo Kimball.***



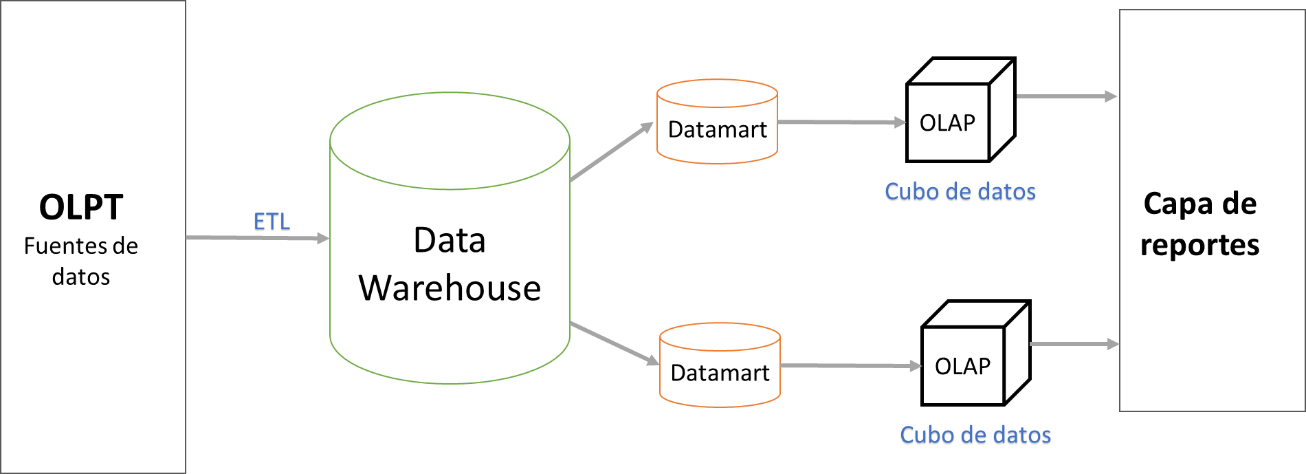
* **Modelo Inmon**

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.

Figura 2.

***Modelo Inmon.***



**3.6 Verificación de valores y escalas**

Los cuadros de mando o D*ashboard*, 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 claras. 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:

* **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.
* **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.
* **Listas**: comúnmente formadas por KPI. En caso de que el cuadro de mando solo esté formado por este tipo de elemento, se denomina *Scorecard*.
* **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 verá 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. Ej: Un gráfico de barras si muestra 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 muestran 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.).
* **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.
* **Alertas visuales:** por lo general, al momento de desarrollar los cuadros de mando se pueden incluir la programación de alertas automáticas con el fin de informar un acontecimiento crítico, ya sea por fechas u otro evento que suceda.
* **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* D*ashboard:* |  |
| --- | --- |

**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 propiamente dicho, pues los motores de bases de datos pueden adaptar además de las consultas estándar SQL, adoptan *script* con lenguajes como: R, Phyton, 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; adicional pueden crear medidas que calcules o cuentes elementos de datos, filas calculadas a partir de datos de otras filas o funciones. | Concepto de almacenamiento digital del sistema informático de base de datos |
| --- | --- |

**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 realizarse una operación que sea el disparador de otras tareas.

Tabla 1.

***Ejemplo de Triggers.***

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

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 presentan 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.

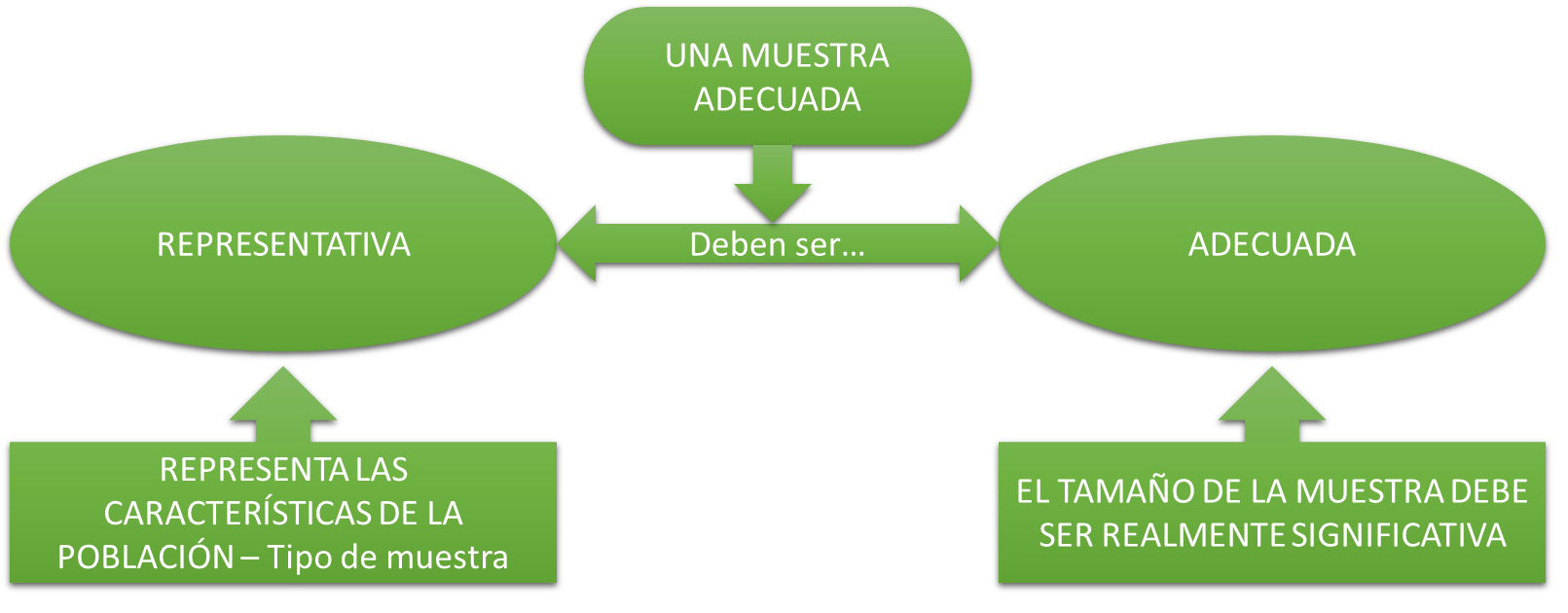
Para la aplicación de estadística inferencial, se tratarán algunos parámetros matemáticos para la aplicación de este modelo estadístico.

**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 utilizamos para recolectar la información aplicando la escala. (Domínguez, 2017). Existen cuatro escalas primarias de medición: nominal, ordinal, intervalo y radio; y algunas escalas más sofisticadas como las escalas multicontenido o multipropósito, en la siguiente tabla se describen cada categoría.

Tabla 2.

***Variables y escalas.***

| VARIABLE | ESCALA | Descripción |
| --- | --- | --- |
| Categórica | Nominal | Clasifica los elementos del conjunto para distribuirlos en grupos. |
| Ordinal | También clasifica elementos, pero además permite hacer escalas de medición comparativa |
| Métrica | Intervalo | * Cuantifica y califica numéricamente los objetos de la categoría. * Permite hacer mediciones simples. * Utiliza escalas continuas * 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. |
| Radio | * Cuantifica y califica numéricamente los objetos de la categoría. * Permite hacer mediciones simples. * Utiliza escalas discontinuas * 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. |

**4.4 Técnicas de análisis estadístico**

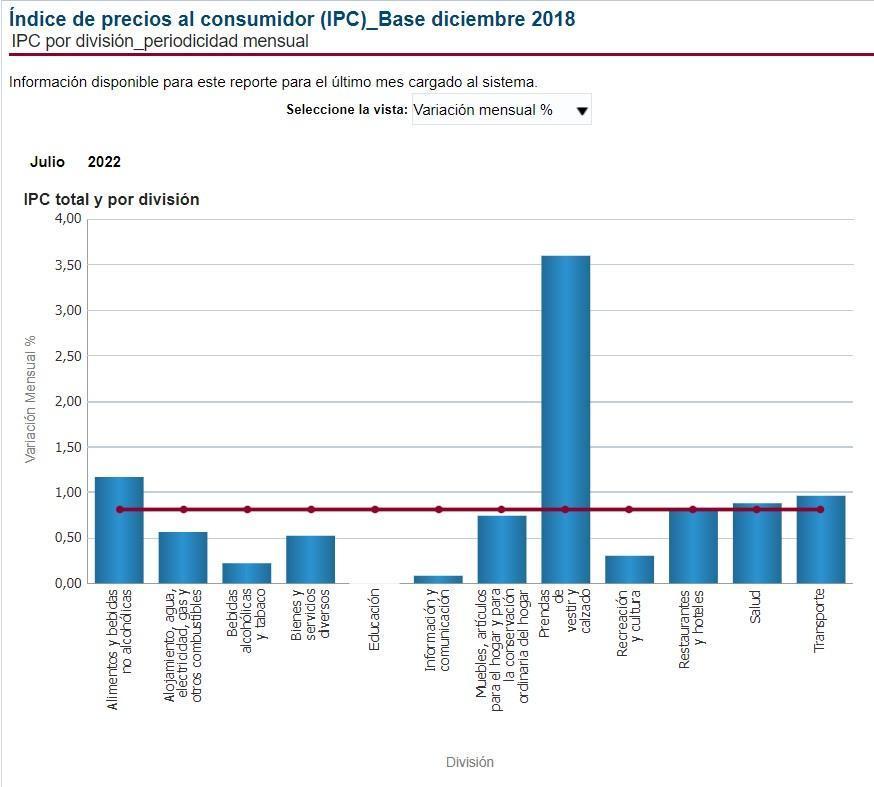
Mencionar parte de las técnicas estadísticas sería un muy extenso y no bastaría un fragmento de un recurso de formación, sin embargo, para efectos de contexto y orientación hacia la analítica de datos y soluciones BI, es fundamental señalar que el análisis estadístico se asocia con las técnicas y prácticas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

Figura 4.

***Índice de precios del consumidor.***



*NOTA: Banco de la República. (2022)*

En la imagen anterior se denota 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.

Con esta información, el negocio podría ir tomando decisiones y estableciendo enfoques comerciales, mirar cómo está su sector en el país y de esa manera ir orientando su análisis de negocio.

Además, existen indicadores propios de las organizaciones, tales como la medición de desempeño de los colaboradores en áreas específicas, validar los datos financieros, los tiempos de procesos, comparar sedes, definir conocimiento específico y general del negocio.

Por lo que la estadística toma un papel muy importante, para ello se mencionará las principales características de la estadística:

Como principio fundamental la estadística contiene los siguientes componentes:



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 en la manera cómo 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.

| Análisis del rendimiento empresarial con gráficos. | Puede ser una tarea dispendiosa pero necesaria antes de emprender procesos técnicos de inteligencia de negocios. |
| --- | --- |

**5.1 Datos univariables**

Son los datos que solo se ocupan de una sola variable, por lo tanto, este tipo de datos es 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 multivariables.**

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

| Su análisis suele ser más robustos 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. | Ilustración del concepto de estadísticas del sitio |
| --- | --- |

**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 ante 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:

* ***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.

***Reglas de negocio derivadas de las políticas empresariales.***



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:

**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 las fuentes para los procesos analíticos, los cuales toman varios programas transaccionales. | Dibujado a mano ilustración de sql de diseño plano |
| --- | --- |

**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 tienes las siguientes características:



**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 video:



**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. |  |
| --- | --- |

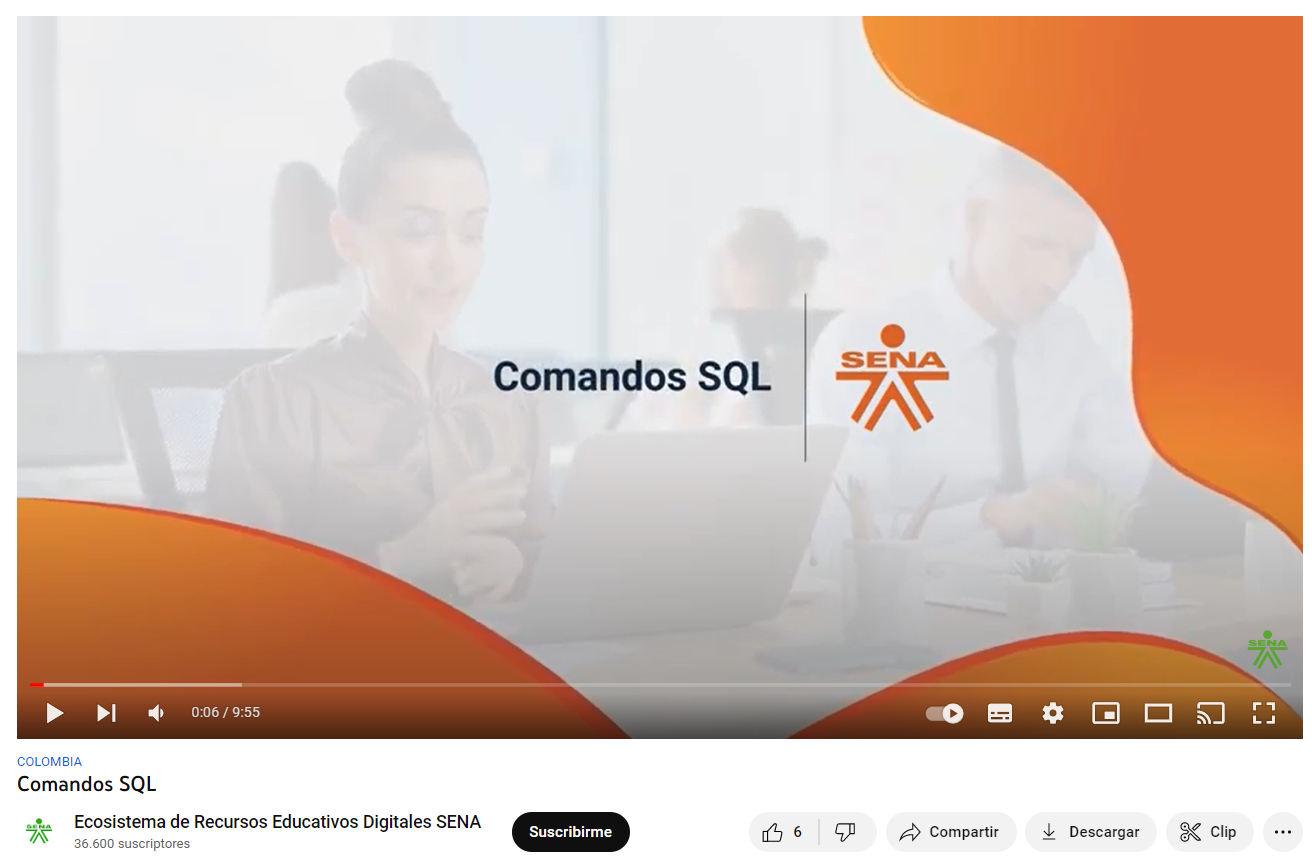
Existen varios motores de bases de datos con núcleo SQL, entre los más conocidos MySQL de uso libre y otra distribución licenciada por Oracle.

En el siguiente video te presentamos las ventajas de Mysql:



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:



**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 JSON 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 e las bases de datos no relacionales o No SQL:



**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:



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, DCL**

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:



**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 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, comprenderás mejor a lo que hace referencia la analítica de datos:



**6.1. Bodega de datos**

La información de las organizaciones, provienen 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. | Ilustración del concepto de datos visuales |
| --- | --- |

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

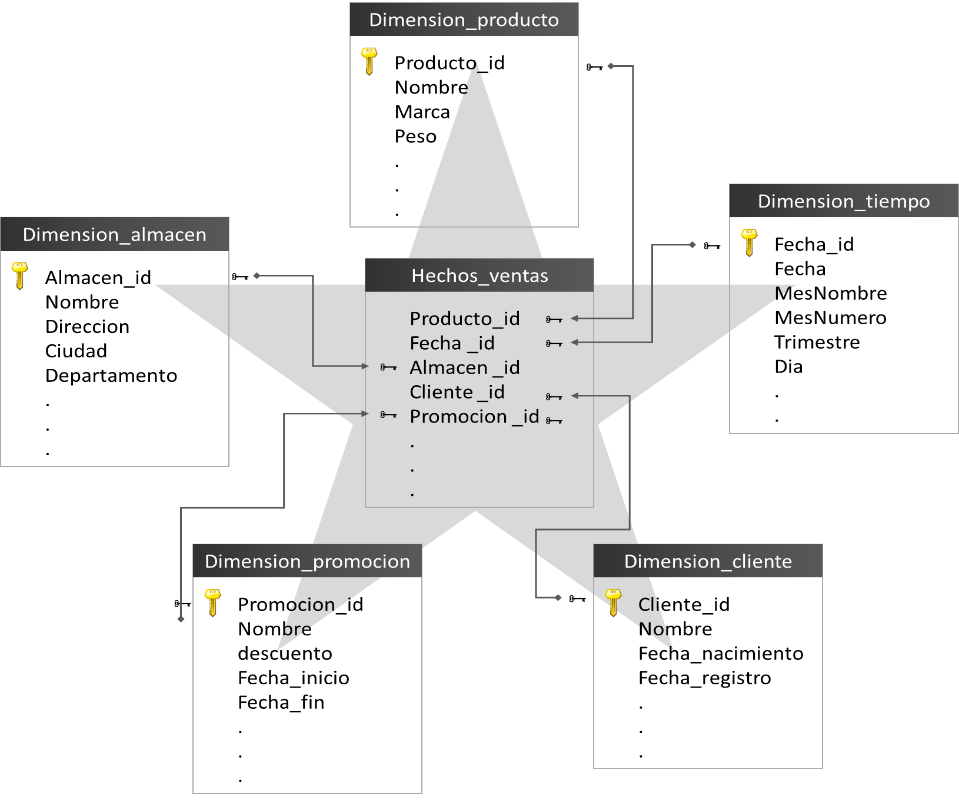
Esta estructura contiene diversas tablas de hechos y dimensiones, que permiten estructurar la información de tal manera que se pueda visualizar de una mejor manera, las tablas que componen la bodega de datos se pueden presentar en diversos diseños, como tipo estrella, copo de nieve, constelaciones, etc.

**6.2 Tipos estrella**

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

Figura 6.

***DWH Tipo Estrella.***



Como se visualiza en la gráfica, muchas tablas giran alrededor de la tabla principal que es la tabla de hechos, que por lo general, contiene las claves primarias de las dimensiones que se asocian, así como las cifras o medidas realizadas previas en el proceso ETL.

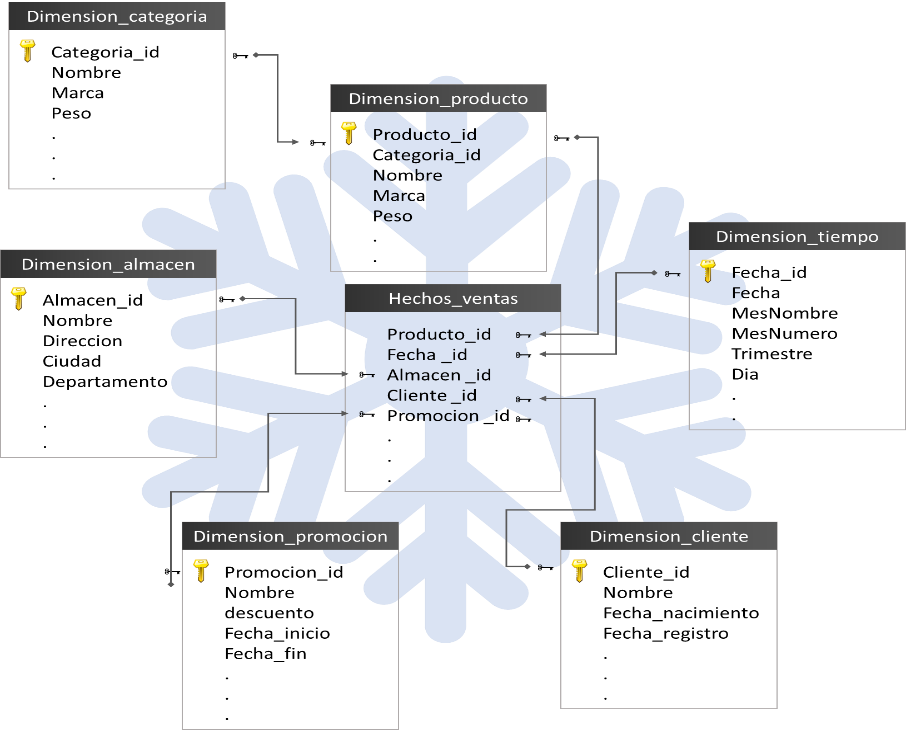
**6.3 Copo de nieve**

La topología copo de nieve o *Snowflake*, se presenta frecuentemente, es 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:

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 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

***Herramientas para el análisis los datos.***

| **Herramientas para el análisis de los datos** | |
| --- | --- |
| **Herramienta** | **Función** |
| **Microsoft Power BI** | 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. |
| **Programación en R** | 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. |
| **SAS** | 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. |
| **Python** | Es una herramienta diseñada para trabajar sobre objetos y procesar datos de forma funcional y estructurada. |
| **Excel** | 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. |
| **Tableau Public** | Esta herramienta de análisis de datos es un *software* gratis que entrelaza las fuentes de información en tiempo real de la Web. |
| **Rapid Miner** | Es una herramienta para realizar análisis predictivos. |
| **Apache Spark** | Es una herramienta que procesa los datos de forma rápida, con algoritmos que le permiten clasificar la información. |
| **Qlik View** | 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 en 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.

Los entornos más eficientes son:



**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 expande 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 como *Code Skulpor* <https://py3.codeskulptor.org/>

**7.3 Librerías**

Las librerías en los lenguajes de programación, son 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 Phyton, 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 bien 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.

1. **SÍNTESIS**

A continuación, se presenta el diagrama que representa el resumen de las temáticas que están desarrolladas en el componente formativo:



1. **ACTIVIDADES DIDÁCTICAS**

| DESCRIPCIÓN DE ACTIVIDAD DIDÁCTICA | |
| --- | --- |
| Nombre de la Actividad | odelamiento, análisis y preparación de datos. |
| Objetivo de la actividad | Identificar los conceptos sobre modelamiento, análisis y preparación de datos con el fin de reforzar los conocimientos adquiridos en el componente. |
| Tipo de actividad sugerida | https://lh6.googleusercontent.com/FvABbrwgrnxeHtlHBpt0FdRjJH5VpWBdLbK0Y-NyHF0MxmgecZxbT_QLjV8v2h_0Gr5ETSV7C0ihbVBtf_VBqiBSiEmxWXfjkqpV-5uTXN3HQPFv2D2_nnl_RHrCosO6f6A5mfAWEenp2HD8W1mNoBi4amrW5VkKn7N1ZGfcpGbi1gj0WLQjzi5eb6uE_Sz7u4uqGA |
| Archivo de la actividad  (Anexo donde se describe la actividad propuesta) | Anexo\_CF13\_ActividadDidactica |

**MATERIAL COMPLEMENTARIO:**

| Tema | Referencia APA del Material | Tipo de material  (Video, capítulo de libro, artículo, otro) | Enlace del Recurso o  Archivo del documento o material |
| --- | --- | --- | --- |
| 1. Modelamiento y gestión de bases de datos | Quintana, G. (2014). Aprende SQL. Universitat Jaume I. Servei de Comunicació i Publicacions. https://elibro-net.bdigital.sena.edu.co/es/lc/senavirtual/titulos/53252 | Libro | <https://elibro-net.bdigital.sena.edu.co/es/lc/senavirtual/titulos/53252> |
| 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. https://youtu.be/KcORNp2A3yg | Video | <https://youtu.be/KcORNp2A3yg> |
| 2.2 Detección de errores y datos faltantes | Codificandobits. (s.f). Mapa paso a paso manejo datos faltantes. Blog. https://www.codificandobits.com/descargas/dl\_202100618\_mapa\_pasos\_manejo\_datos\_faltantes.pdf | Guía | <https://www.codificandobits.com/descargas/dl_202100618_mapa_pasos_manejo_datos_faltantes.pdf> |
| 2.3 Identificación de variables importantes | Sotaquirá, M. (2021). Guía completa para el manejo de datos faltantes. Blog. https://www.codificandobits.com/blog/manejo-datos-faltantes/ | Artículo | <https://www.codificandobits.com/blog/manejo-datos-faltantes/> |
| 3.7 Procedimientos almacenados y funciones | Calbimonte, D. (2019). Funciones frente a los procedimientos almacenados en SQL Server. https://www.sqlshack.com/es/funciones-frente-a-los-procedimientos-almacenados-en-sql-server/ | Tutorial | <https://www.sqlshack.com/es/funciones-frente-a-los-procedimientos-almacenados-en-sql-server/> |
| 4. Análisis exploratorio de datos | Codificandobits. (s.f). Guía paso a paso análisis exploratorio. Blog. https://www.codificandobits.com/descargas/dl\_202100611\_mapa\_pasos\_analisis\_exploratorio.pdf | Mapa conceptual | <https://www.codificandobits.com/descargas/dl_202100611_mapa_pasos_analisis_exploratorio.pdf> |
| 5.8 Algebra relacional | Cidecam. (2021). Algebra Relacional. Página web.  http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro14/33\_algebra\_relacional.html | Artículo | <http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro14/33_algebra_relacional.html> |
| 7.3 Librerías | Ecosistema de Recursos Educativos Digitales SENA. (2021). Manejo de datos con Pandas. SENA. https://youtu.be/P6LmIUMJg3U | Vídeo | <https://youtu.be/P6LmIUMJg3U> |

1. **GLOSARIO:**

| **TÉRMINO** | **SIGNIFICADO** |
| --- | --- |
| **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. |
| **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. |
| **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. |
| **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. |
| **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 |
| **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. |
| **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. |
| **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. |

1. **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>

Conza, A. (15 de abril de 2018). ESTÁNDARES PARA DISEÑO E IMPLEMENTACIÓN DE BASES DE DATOS RELACIONALES. Obtenido de http://adrianconza.com/: <http://adrianconza.com/estandares-de-bases-de-datos/>

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

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

Domínguez, S. (1 de agosto de 2017). Definición y Ejemplos de las Escalas de Medición. Obtenido de Stella Domínguez: <https://stelladominguezcom.wordpress.com/2017/08/01/escalas/>

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. Obtenido de MyCloudWiki: <https://www.mycloudwiki.com/san/data-and-information-basics/>

Pang, A., Markovski, M., & Ristik, M. (22 de septiembre de 2022). Los 10 principales proveedores de software de análisis y BI, tamaño del mercado y pronóstico del mercado 2021-2026. Obtenido de 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. Obtenido de <https://www-ebooks7-24-com.bdigital.sena.edu.co/?il=9094>

1. **CONTROL DEL DOCUMENTO**

|  | **Nombre** | **Cargo** | **Dependencia** | **Fecha** |
| --- | --- | --- | --- | --- |
| Autor (es) | Jaime Hernán Tejada | Experto Temático | Centro de la Industria, la empresa y los servicios - CIES | Noviembre de 2022 |
| Giovanna Andrea Escobar Ospina | Diseñador Instruccional | Sena - Norte de Santander Centro CIES | Noviembre de 2022 |
| Silvia Milena Sequeda Cárdenas | Asesora metodológica y pedagógica | Regional Distrito Capital -Centro de Diseño y Metrología. | Noviembre de 2022 |

1. **CONTROL DE CAMBIOS**

**(Diligenciar únicamente si realiza ajustes a la Unidad Temática)**

|  | Nombre | Cargo | Dependencia | Fecha | Razón del Cambio |
| --- | --- | --- | --- | --- | --- |
| Autor (es) |  |  |  |  |  |

**Nota:**Para la propuesta instruccional se deben tener en cuenta las métricas desarrolladas en el equipo:

<https://drive.google.com/drive/u/1/folders/1UiJvaklSCICR4BaQ7ga_q04JFa53h_u_>