

DATAWAREHOUSE Y MINERIA DE DATOS

Ciclo 2-2020



1

Análisis Multidimensional



Introducción



- ▶ Desde el inicio de la era de la computación las organizaciones han usado los datos desde sus sistemas operacionales para **atender sus necesidades de información.**
- ▶ Algunas proporcionan acceso directo a la información contenida dentro de las aplicaciones operacionales.

Introducción



- ▶ Otras, han extraído los datos desde sus bases de datos operacionales para combinarlos de varias **formas no estructuradas**, en su intento por atender a los usuarios en sus necesidades de información.
- ▶ Ambos métodos han evolucionado a través del tiempo y ahora las organizaciones manejan datos aislados e incompletos, sobre los cuales, en la mayoría de las veces, **se toman importantes decisiones**.

¿Qué es el análisis multidimensional?



- ▶ El análisis multidimensional es un recurso que le permite **evaluar los datos de su empresa desde diferentes puntos de vista**.
- ▶ Por esta razón, la información está **organizada por jerarquía**, para facilitar el estudio de estos datos. Como tal, entendemos que una jerarquía es parte de los diferentes niveles que conforman una dimensión, que no es más que una unidad de análisis.

¿Cómo puede el análisis multidimensional ayudar a la empresa?

EJEMPLO

Dimensión: Región

Para que un gerente de una empresa cree estrategias eficientes que realmente atraigan al **público objetivo**, es necesario analizar varias características que conforman el perfil del cliente, como su **región**.

La región se considera una dimensión cuyos niveles jerárquicos están divididos (y organizados) por países, estados, ciudades y áreas de operación.



¿Cómo puede el análisis multidimensional ayudar a la empresa?

Por lo tanto, en una empresa, hay varias dimensiones (como **productos, tiempo, clientes, producción, etc.**) que no solo pueden, sino que deben analizarse juntas, mediante análisis multidimensional, para obtener mejores resultados.

Después de todo, el análisis multidimensional permite al administrador cruzar una o más dimensiones. Un claro ejemplo de esto es el **desarrollo de un producto que satisfaga las necesidades del público en una región determinada** y, después de su lanzamiento, la posibilidad de analizar el tiempo que lleva venderlo.



En otras palabras, al verificar la forma en que el **análisis multidimensional** organiza los datos, el gerente puede **identificar qué productos se vendieron más en un período determinado, en qué región tuvo la mejor salida y qué canal de ventas se utilizó para generar este retorno.**



¿Cómo se puede usar el análisis multidimensional en los costos de la empresa?

- El uso del análisis multidimensional es esencial para **garantizar una gestión de costos más eficiente**.
- Esto se debe a que ayudará a determinar la rentabilidad de la empresa, ya que tiene en cuenta todos sus gastos y costos. Es decir, cruza dimensiones relacionadas con el producto, el cliente, la demanda, la región, etc.



¿Cómo se puede usar el análisis multidimensional en los costos de la empresa?

- A través de esta acción, tiene en sus manos un informe completo y preciso sobre la situación financiera real de la compañía, con la ventaja de poder identificar los gastos que pueden eliminarse y, sobre todo, reducir otros gastos sin tener que renunciar a la calidad y la eficiencia.

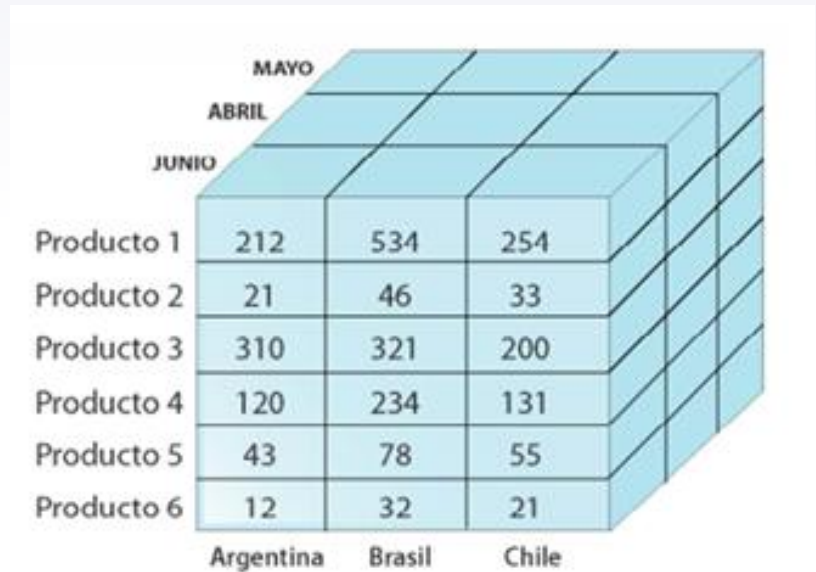


¿Cómo hacer un análisis de costos multidimensional?

- ▶ El análisis multidimensional es una herramienta que funciona de manera muy similar a una tabla dinámica, en la que ingresa la información que será útil para el análisis. Por esta razón, es extremadamente importante **definir las dimensiones de acuerdo con las características de su propio negocio.**
- ▶ Por lo tanto, es posible realizar análisis multidimensionales mediante procesamiento analítico en línea (OLAP).

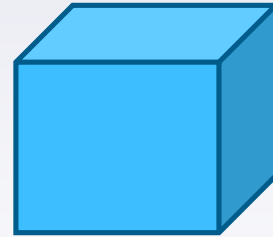


2 Cubos OLAP



	JUNIO	ABRIL	MAYO
Producto 1	212	534	254
Producto 2	21	46	33
Producto 3	310	321	200
Producto 4	120	234	131
Producto 5	43	78	55
Producto 6	12	32	21
	Argentina	Brasil	Chile

Cubos OLAP



- ▶ En el mundo de las soluciones para **Business Intelligence**, una de las herramientas más utilizadas por las empresas son las **aplicaciones OLAP** (OLAP significa "On-Line Analytical Processing"), ya que las mismas han sido creadas en función a **bases de datos multidimensionales**, que permiten procesar grandes volúmenes de información, en campos bien definidos, y con un acceso inmediato a los datos para **su consulta y posterior análisis**.

Historia

- ▶ El **término Cubo OLAP**, que nació de la mano de Edgar F. Codd, de la compañía EF Codd & Associates, sólo se utilizaban bases de datos relacionales para el proceso de la información, con sistemas tales como el ROLAP.

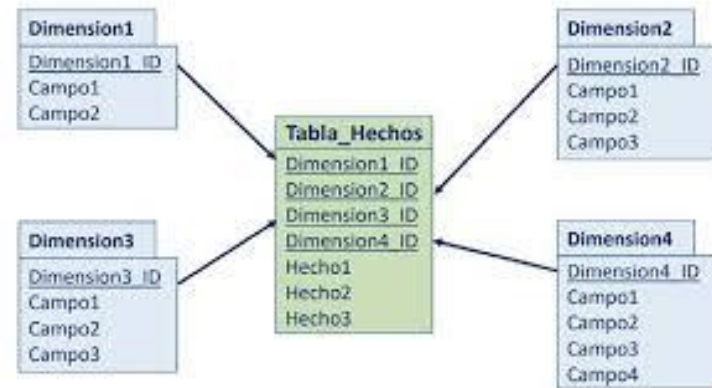
Datawarehouse

Cubes OLAP



Historia

- ▶ Gracias a la incorporación de las bases de datos de tipo multidimensional, y el nacimiento del nuevo concepto Cubo OLAP, las herramientas de soluciones para sistemas Business Intelligence han avanzado notablemente en cuanto a las prestaciones que estas aplicaciones brindan a las empresas, donde la información confiable, precisa y en el momento oportuno, son uno de los bienes más preciados



► Funciones

- ▶ Las herramientas OLAP proporcionan a las compañías un sistema confiable para procesar datos que luego serán utilizados para llevar a cabo **análisis e informes que permitan mejorar las operaciones productivas, tomar decisiones inteligentes y optimizar la competitividad en el mercado.**
- ▶ Para funcionar, las aplicaciones OLAP utilizan un tipo de base de datos que posee la característica de ser multidimensional, denominada comúnmente **Cubo OLAP.**

► Funciones

- ▶ Básicamente, el Cubo OLAP, es una base de datos que **posee diversas dimensiones**, ampliando las posibilidades que hasta el momento ofrecían las conocidas hojas de cálculo.
- ▶ Mediante la incorporación de estos cubos, se han ampliado las posibilidades de las bases de datos relacionales, permitiendo el procesamiento de **importantes volúmenes de información**, que de lo contrario sería imposible realizar.

► Funciones

- ▶ Cada una de las dimensiones que posee la base de datos **incorpora un campo determinado para un tipo de dato específico**, que luego podrá ser comparado con la información contenida en el resto de dimensiones, para hacer posible la evaluación y posteriores informes de la información realmente relevante para una compañía.
- ▶ Una base de datos multidimensional puede contener **varios cubos** que extenderán las posibilidades del sistema OLAP con el cual se trabaja.

Características

- ▶ En este **modelo** los datos son vistos como cubos los cuales consisten en **categoría descriptivas** (dimensiones) y **valores cuantitativos** (medidas).
- ▶ Es un modelo multidimensional de datos, que simplifica a los usuarios realizar **consultas complejas**, arreglar datos en un reporte, cambiar de datos resumidos a datos detallados, etc.
- ▶ Los Cubos OLAP son vectores en los cuales se dispone la información, y gracias a la ordenada jerarquía que posee es posible llevar a cabo un análisis rápido de los datos.

Operaciones Analíticas Básicas

- ▶ **Consolidación:**

- ▶ Este comprende el conjunto de datos.
- ▶ Esto puede involucrar acumulaciones simples o agrupaciones complejas que incluyen datos interrelacionados.



Operaciones Analíticas Básicas

- ▶ **Drill-Down:**

- ▶ OLAP puede moverse en dirección contraria y presentar automáticamente datos detallados que abarcan datos consolidados.

- ▶ **Slicing-Dicing:**

- ▶ Se refiere a la capacidad de visualizar a la base de datos desde diferentes puntos de vistas.

► Tipos de arquitecturas OLAP

MOLAP

- ▶ Utiliza BDD Multidimensionales
- ▶ Tiene una mayor implantación en el análisis multidimensional
- ▶ La arquitectura consta de dos niveles, BDD multidimensional y motor analítico
- ▶ El nivel de aplicación se encarga de los requerimientos OLAP
- ▶ Nivel de presentación se integra con el motor proporcionado la interfaz de usuario.

Tipos de arquitecturas OLAP

MOLAP

► Proceso

- La información de los sistemas operacionales se cargan en MOLAP, mediante rutinas por paquetes o lotes.
- Una vez cargada la información en la base de datos multidimensional, estos son analizados mediante paquetes o lotes.
- Se generan algoritmos e índices de tablas hash para que las consultas se ejecuten con mayor rapidez.
- El usuario realiza las consultas a través de la interfaz.

Tipos de arquitecturas OLAP

ROLAP

- ▶ La arquitectura accede a los datos de Datawarehouse para proporcionar un análisis OLAP.
- ▶ La arquitectura consta de tres niveles: Una BDD relacional que maneja los requerimientos de almacenamiento, un motor ROLAP que proporciona la funcionalidad analítica, y BDD relacionales para el manejo, acceso y obtención de los datos.
- ▶ El motor ROLAP, se integra con los niveles de presentación a través de los cuales los usuarios realizan las consultas.

Tipos de arquitecturas OLAP

ROLAP

► Proceso

- Carga de datos desde sistemas operacionales en Datawarehouse.
- Se ejecutan rutinas desde la BDD para agregar los datos.
- Se crean índices para optimizar los tiempos de acceso a las consultas.
- Los usuarios realizan las consultas mediante los motores ROLAP, que se transforman dinámicamente a consultas SQL
- Las consultas se ejecutan en las BDD relacionales y se ejecutan mediante tablas cruzadas.

Tipos de arquitecturas OLAP

▶ **HOLAP**

- ▶ Se presenta como una solución híbrida entre ROLAP y MOLAP, conservando los aspectos positivos de cada una de las arquitecturas.
- ▶ Al igual que las dos arquitecturas anteriores, utiliza cubos multidimensionales que brindan una mayor rapidez en las consultas.
- ▶ No hay unas claras especificaciones de cómo debe ser una arquitectura HOLAP en el mercado, en lo que si se coincide es que debe estar dividida en tablas relacionales y almacenamientos especializados.

► Tipos de arquitecturas OLAP

► **HOLAP**

- ▶ Normalmente, las tablas relacionales son utilizadas para las grandes cantidades de datos específicos, mientras que para pequeñas cantidades o datos menos específicos, utiliza almacenamientos especializados.
- ▶ Las herramientas HOLAP, pueden utilizar tanto cubos pre-calculados como tablas relacionales generando datos dinámicamente.

► Arquitectura y creación de un cubo

- Para la creación de un cubo se deben seguir 3 pasos:
 1. El primer paso es la definición del cubo
 2. Especificar la estrategia de Suma de datos diseñando las agregaciones (elementos pre-calculados de datos)
 3. Y el último paso es la carga del cubo para procesarlo.
- Para definir un cubo se selecciona la tabla objetivo y también las medidas dentro de esa tabla, luego se selecciona las dimensiones compuesta de una o más columnas de otra tabla.

Se tomará como base para la realización del ejemplo, el siguiente esquema en estrella:



Como primer paso se creará un cubo multidimensional llamado "Cubo de Ventas", gráficamente:



Cubo de Ventas

Luego se crearán dos atributos:

De la tabla de dimensión “PRODUCTOS”, se tomará el campo “Producto” para la creación del atributo denominado:

- **“PRODUCTOS - Producto”**.

De la tabla dimensión “MARCAS”, se tomará el campo “Marca” para la creación del atributo denominado:

- **“MARCAS - Marca”**.



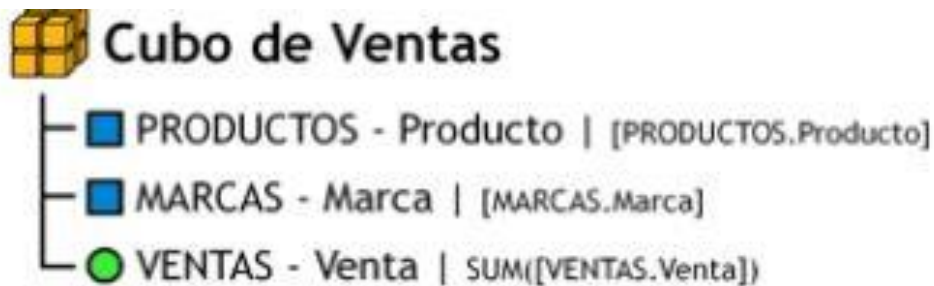
También se creará un indicador:

De la tabla de hechos “VENTAS”, se calculara la suma del hecho “Venta” para crear el indicador denominado:

- **“VENTAS - Venta”.**

La fórmula utilizada para crear este indicador es la siguiente:

- **“VENTAS - Venta” = SUM(VENTAS.Venta).**



En este momento, tenemos un cubo multidimensional de dos dimensiones, cuya representación matricial sería la siguiente:

PRODUCTOS - Producto	P1	40	25	60
	P2	21	55	45
	P3	13	32	43
		M1	M2	M3
		MARCAS - Marca		

Este cubo posee dos ejes o dimensiones, “PRODUCTOS - Producto” y “MARCAS - Marca”. La intersección de los ejes representa las ventas de cada producto con su respectiva marca (indicador “VENTAS - Venta”).

Por ejemplo:

- Las ventas asociadas al producto “P1” y a la marca “M1” son 40.
- Las ventas asociadas al producto “P1” y a la marca “M2” son 25.
- Las ventas asociadas al producto “P1” y a la marca “M3” son 60.

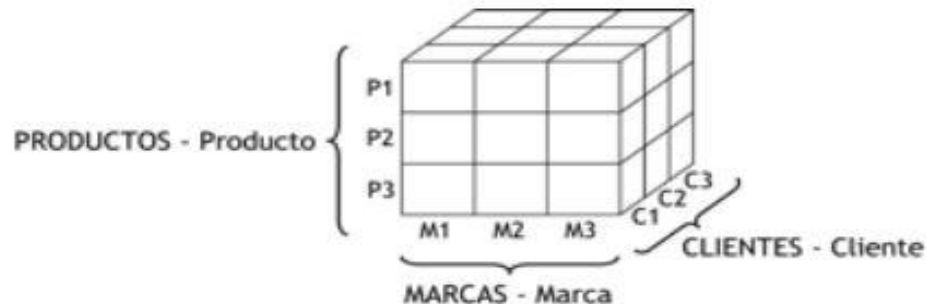
Ahora, al cubo planteado se le agregará un nuevo atributo:

De la tabla de dimensión “CLIENTES”, se tomará el campo “Cliente” para la creación del atributo denominado:

- “**CLIENTES - Cliente**”.



De esta manera, ahora tenemos un cubo multidimensional de tres dimensiones, cuya representación matricial sería la siguiente:



En este caso los valores del indicador “VENTAS - Venta” están dados de acuerdo a las ventas de cada producto, de cada marca, a cada cliente.

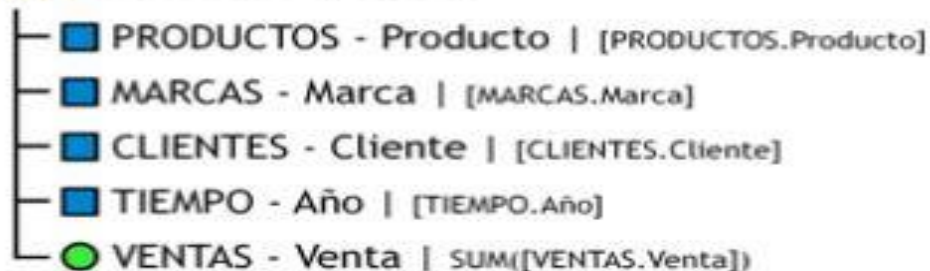
Para finalizar, se añadirá un cuarto atributo al cubo:

De la tabla de dimensión “TIEMPO”, se tomará el campo “Año” para la creación del atributo denominado:

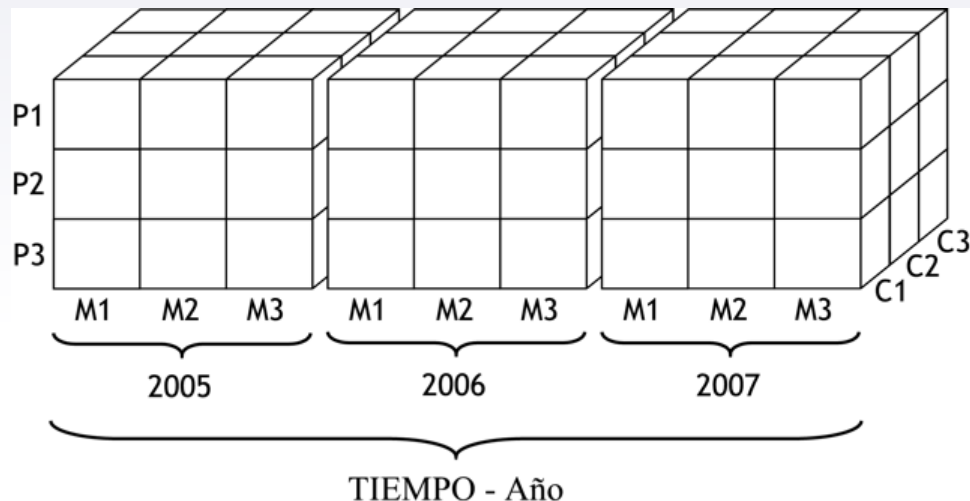
- **“TIEMPO - Año”**.



Cubo de Ventas



Entonces, la representación matricial del cubo multidimensional resultante sería la siguiente:



Los valores del indicador “VENTAS - Venta” en este momento, estarán condicionados por las ventas de cada producto, de cada marca, a cada cliente, en cada año.

Esta última imagen, demuestra claramente los conceptos de la tabla de dimensión tiempo, donde pueden existir diferentes versiones de la situación del negocio.

Cabe aclarar que pueden crearse tantos cubos se deseen y que los mismos pueden coexistir sin ningún inconveniente.

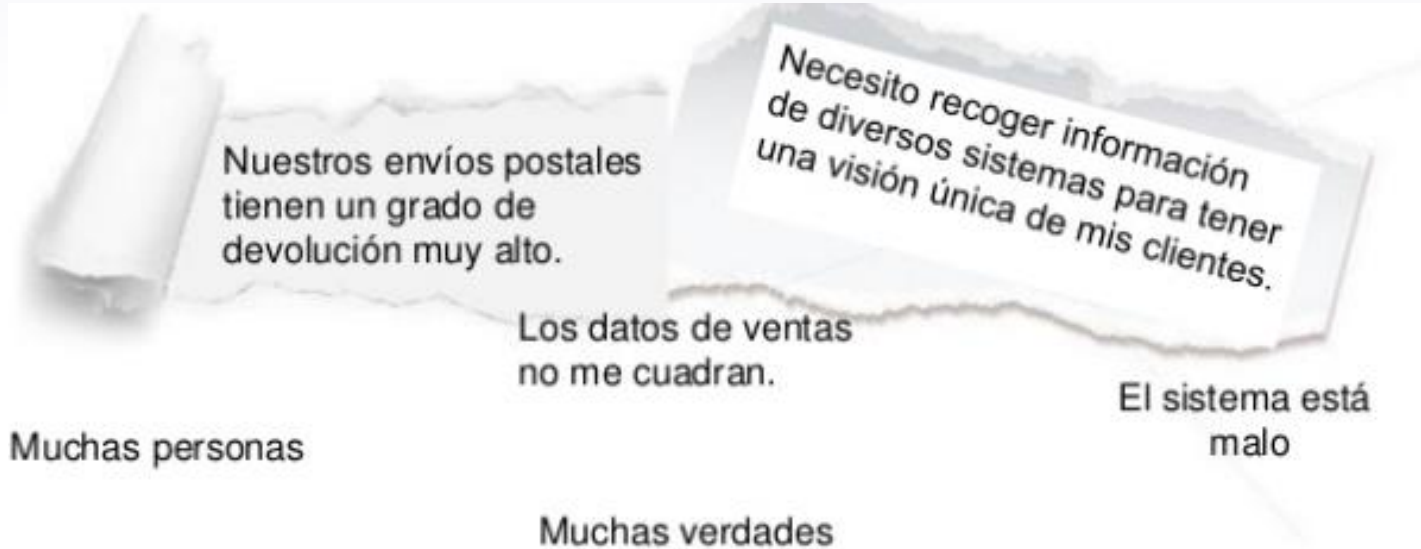
3

Calidad de datos



¿Qué es Calidad de Datos?

- ▶ Algunos problemas



¿Qué es Calidad de Datos?

- ▶ **Calidad de datos** es la cualidad de un conjunto de información recogida en una base de datos, un sistema de información o un DataWarehouse que reúne entre sus atributos la **exactitud, completitud, integridad, actualización, coherencia, relevancia, accesibilidad y confiabilidad** necesarias para resultar útiles al procesamiento, análisis y cualquier otro fin que un usuario quiera darles.



¿ Por qué necesito calidad de datos?

- ▶ Imprescindible para una toma de decisiones correcta.
- ▶ Pieza básica en la gestión de información corporativa.
- ▶ Impacto directo en el negocio.
- ▶ Retorno de la inversión inmediato.
 - ▶ “Si los datos son la materia prima con la cual creamos la Información; entonces la confiabilidad de la información depende directamente de la calidad de los datos utilizados para producirla”



Ciclo de vida en la calidad de los datos



Calidad de Datos en BI

- ▶ **La calidad del dato** en Business Intelligence (BI) es una de los **principales factores de éxito**.
- ▶ En el mundo de la analítica existe una ley universal: si la calidad de los datos es pobre, los informes que se hacen con ellos también lo son. Por ende, las decisiones tomadas en base a ellos tampoco suelen ser acertadas.



► Calidad de Datos en BI

- ▶ De una forma u otra, todas las empresas tienen que lidiar con la falta de calidad de los datos en algún momento. Sin embargo, muchas aún desconocen que, para corregir sus principales problemas, **lo más esencial es tener una adecuada arquitectura de datos, organizada y relevante.**
- ▶ Partiendo de esta premisa, llevar a cabo mejoras y actualizaciones continuas que permitan la **explotación de los datos** con fines comerciales o para su uso en la toma de decisiones es vital en cualquier organización.

► Calidad de Datos en BI

- ▶ La calidad del dato en Business Intelligence **garantiza una menor cantidad de obstáculos para tomar decisiones.**
- ▶ De hecho, los datos de calidad facilitan el conocimiento de la realidad. Por ello, es la piedra angular sobre la cual basar decisiones de todo tipo, desde las cotidianas hasta las que requieran de análisis más profundos.

Características

- ▶ Para que los datos sean de calidad, normalmente se busca que posean de algunas características que determinan que son aplicables y útiles, que estén contextualizados y sean relevantes.
- ▶ Así la calidad del dato en Business Intelligence viene determinada por una serie de aspectos:
 - ▶ **Mejor si son numerosos.** En la mayoría de las ocasiones se tienen más datos de los que se necesitan. Sin embargo, es mejor que sobren y no que falten. En ocasiones te puedes enfrentar a situaciones imprevistas y, mientras más tengas, más fácil te resultará hallar soluciones.

Características

- ▶ **Tienen en cuenta la historicidad.** La mayoría de los informes suelen ser muy actuales. Sin embargo, siempre es recomendable tener información histórica para comprobar la tendencia o el crecimiento de periodos más amplios. Así, es más fácil prever posibles problemas y solventarlos.
- ▶ **Precisos y completos.** Para que los datos sean de calidad deben ser limpios y veraces. Sin esto, las decisiones tomadas basándose en ellos serán defectuosas.
- ▶ **Expresados de tal forma que todo el mundo pueda comprenderlos.** Si expresas los datos en ID o códigos generados por el sistema de forma automática es muy posible que sean incomprensibles para la gran mayoría.

Tendencias en datos y BI para 2020

- ▶ Que la calidad de los datos en Business Intelligence es la tendencia más relevante de 2020 es una realidad.
- ▶ Pero, ¿cuáles son las principales tendencias de BI en 2020 que permitirán garantizar una mayor calidad de la información?
 - 1. Ser una empresa que recoge datos en tiempo real ya no es una opción:** Hoy en día es vital para las empresas recibir y enviar datos en tiempo real y al lugar correcto.
 - 2. El gran papel del Big Data:** Con formatos y fuentes de datos más variados y fragmentados que nunca.

Tendencias en datos y BI para 2020

- 3. La tecnología asociativa:** permite una mayor flexibilidad en las consultas y la obtención de respuestas de forma más rápida ya que los los datos están almacenados en una memoria RAM que es donde se forman las asociaciones y recibe el nombre de “nube de datos”.
- 4. Los catálogos de datos:** Los datos que recopilan las empresas tienen una gran multitud de fuentes y una distribución fragmentada.

Tendencias en datos y BI para 2020

- 5. **La inteligencia artificial:** permite que a medida que más personas se conectan en el mundo resulte más sencillo que los sistemas analicen y cataloguen información
- 6. **El procesamiento del lenguaje natural mejorado:** La IA procesa con precisión el lenguaje y lo modela de una manera similar a la forma en que los humanos leen.
- 7. **El Business Intelligence móvil:** puede ayudar a que todos, especialmente los tele trabajadores, tengan acceso a los datos en cualquier momento y lugar.

Fuentes de consultas

- ▶ <http://tecnomark.es/analisis-multidimensional->
- ▶ <https://www.myabcm.com/es/blog-post/como-la-analisis-multidimensional-ayuda/>
- ▶ https://www.ecured.cu/Cubos_OLAP
- ▶ <https://www.acceso360.com/la-calidad-del-dato-en-business-intelligence/#:~:text=La%20calidad%20del%20dato%20en,ellos%20tampoco%20suelen%20ser%20acertadas.>

Gracias!

¿Preguntas?

