
Universidad
Tecmilenio®

Universidad Tecmilenio

Fundamentos de big data

# Tema 7.
# Bases de datos analíticas.

Universidad Tecmilenio

**Cuando se maneja un gran volumen de datos operacional, llega el momento de decidir en dónde se van a almacenar estos datos.**

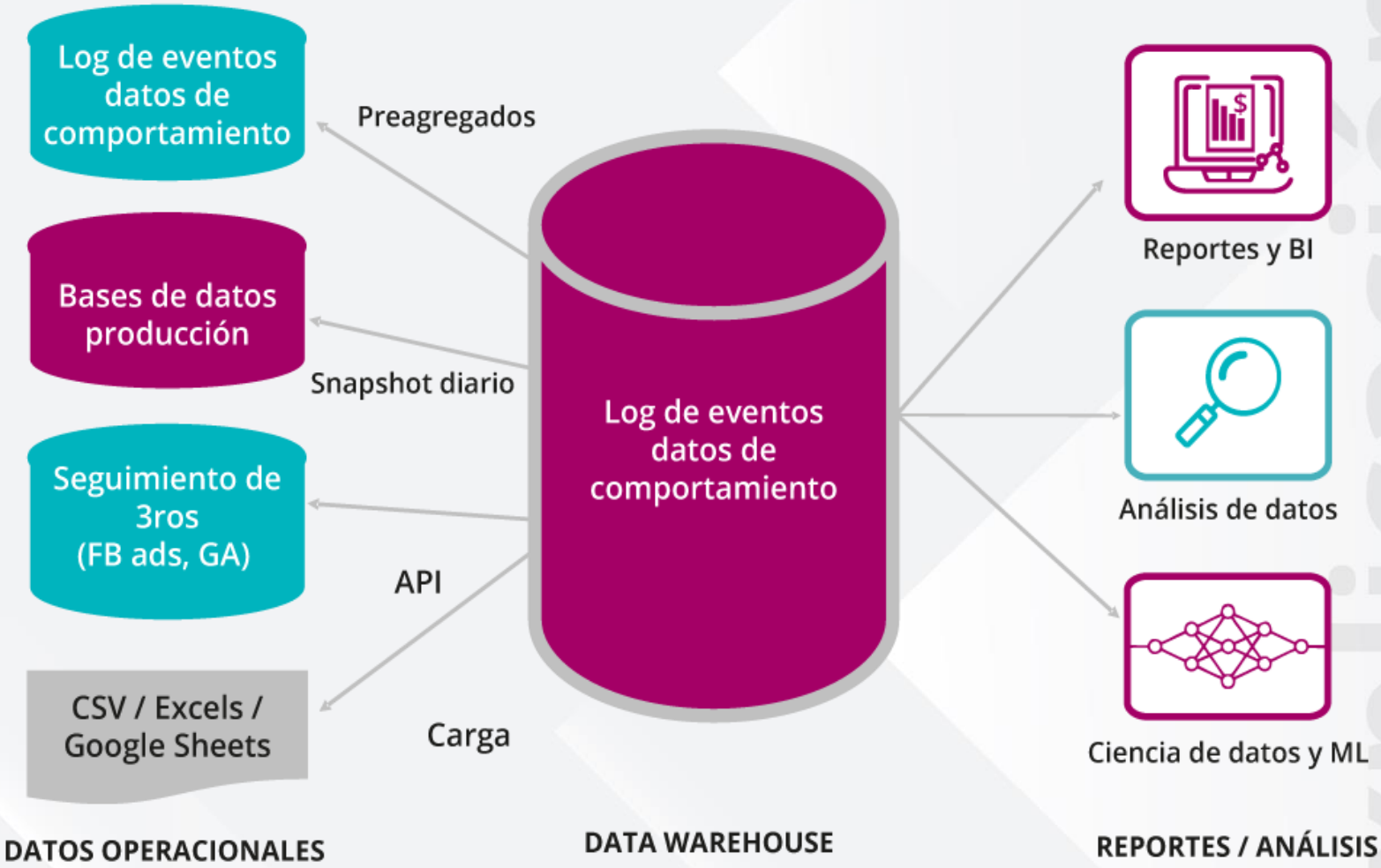
En ese momento es cuando se necesita elegir una **base de datos especializada en la analítica**, que permite hacer búsquedas más rápidas y eficientes, que permita organizar mejor la información, que incluso sea óptima en tamaño y espacio usado, además de que permita almacenar todo el tipo de datos que la empresa requiere de acuerdo con las políticas fijadas para ello.

Universidad Tecmilenio

Una **base de datos** analítica almacena y administra big data, incluidos datos comerciales, de mercado y de clientes para el análisis de inteligencia comercial (BI).

Están optimizadas para proporcionar tiempos de **respuesta** de consulta **rápidos** y análisis **avanzados**. También son más escalables que las bases de datos tradicionales y, a menudo, son bases de datos en columnas.

Figura 1. Configuración común para base de datos de analítica.




Fuente: Holistics. (2021). *The analytics setup guidebook*. Recuperado de https://www.holistics.io/books/setup-analytics/a-modern-analytics-stack/

Universidad Tecmilenio

Existen varias fuentes de las que se alimenta la base de datos de analítica para una empresa u organización.

- **Log de eventos.**
- **Bases de datos de producción.**
- **Seguimiento de servicios de terceros.**
- **CSV, excels, google sheets.**

Universidad Tecmilenio

Esta información que se genera puede derivar en las siguientes salidas o resultados:

- **Reportes o BI:** Existen **informes de base** con los que el negocio pudiera conocer su salud financiera, productividad, alcance de metas y resultados.
- **Análisis de datos**. para poder ofrecer más conocimientos de la propio empresa.
- **Ciencia de datos y/o ML.** Posibilita generar la localización de una nueva sucursal, o bien hacer modelos o automatizaciones AI.

Explicación

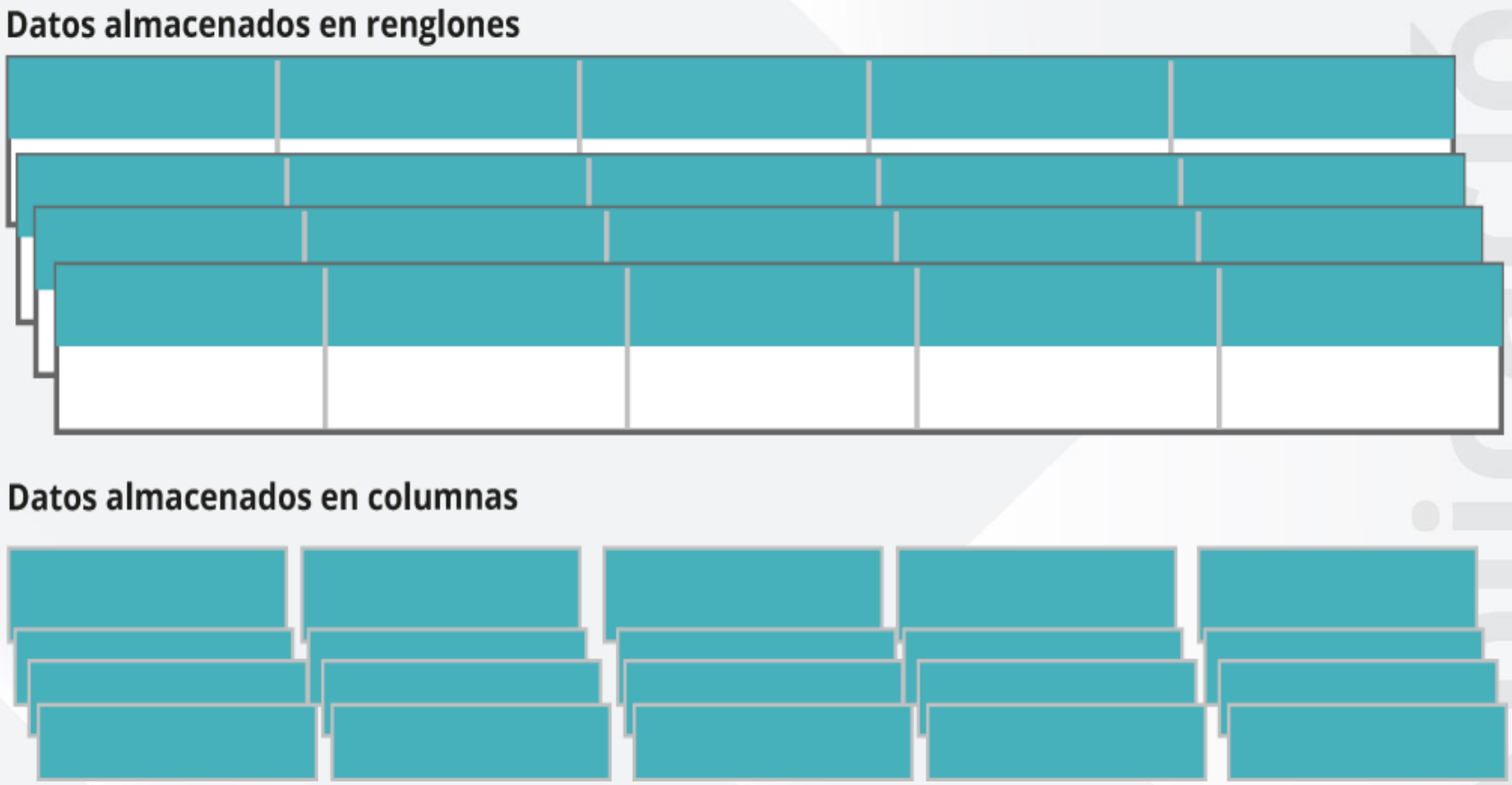
## Tipos de bases de datos de analítica en el mercado.

Las características de las bases de datos analíticas incluyen almacenamiento basado en columnas, carga en memoria de datos comprimidos y la capacidad de buscar datos a través de múltiples atributos.

Actualmente existen cinco tipos principales en el mercado:

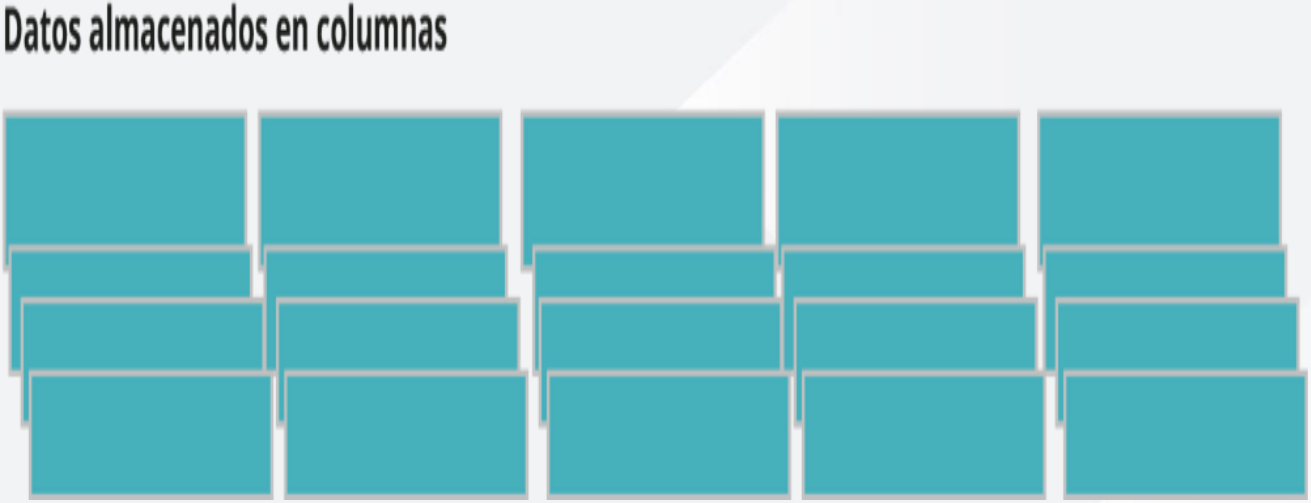
1. Bases de datos en columnas
2. Almacenes de datos
3. Bases de datos en memoria
4. Bases de datos de procesamiento paralelo masivo (MPP)
5. Bases de datos de procesamiento analítico en línea (OLAP)

Figura 2. Diferencia entre base de datos de renglones y una de columnas.




Fuente: Scylla. (2022). *Columnar Database.* Recuperado de https://www.scylladb.com/glossary/columnar-database/

Universidad Tecmilenio




Organizan los datos por columnas en lugar de filas, lo que reduce la cantidad de elementos de datos que debe leer el motor de la base de datos.

**Característica principal:** Optimizan datos para funciones y operaciones agregadas en columnas de datos, almacenando valores de datos de forma contigua, con el mismo tipo de datos y significado semántico.

**Ejemplos:**

- Apache Cassandra
- MariaDB
- DataStax
- Microsoft Azure Cosmos DB
- ScyllaDB

Figura 3. Almacén de datos o Data Warehouse


Fuente: DataChannel. (2021). *What is data warehousing? How it works, types, and general stages*. Recuperado de https://datachannel.co/blogs/introduction-to-data-warehousing/

Combinan la base de datos con hardware y herramientas de inteligencia comercial en una plataforma integrada.

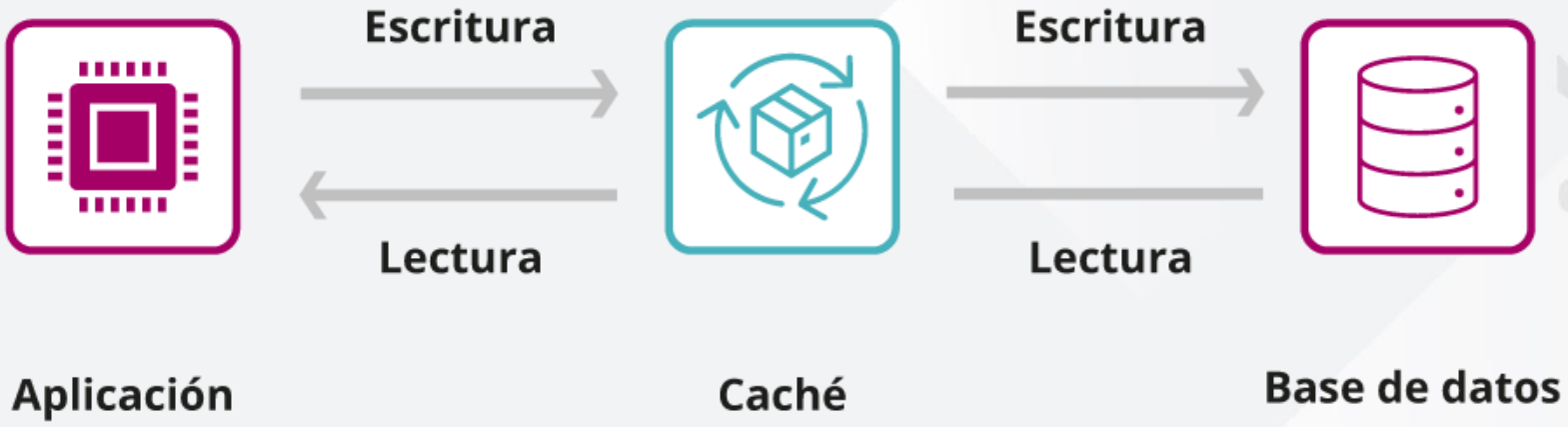
**Existen tres tipos:**

1. Almacén de datos empresarial (Entrerprice Data Warehouse)
2. Almacén de datos operativos (ODS)
3. Cubo de datos (Data Mart)

**Ejemplos de Dataware House :**

- Apache Hadoop
- IBM Netezza Performance Server
- Oracle Exadata
- Snowflake
- Vertica

Figura 4. Base de datos en memoria (caché).




Fuente: GridGain. (s.f.). *In-memory cache*. Recuperado de https://www.gridgain.com/resources/glossary/in-memory-computing-platform/in-memory-cache

Cargan los datos de origen en la memoria del sistema en un formato comprimido y no relacional para agilizar el trabajo involucrado en el procesamiento de consultas.

**Ejemplos:**
Redis
SAP HANA Database
Tarantool
VoltDB

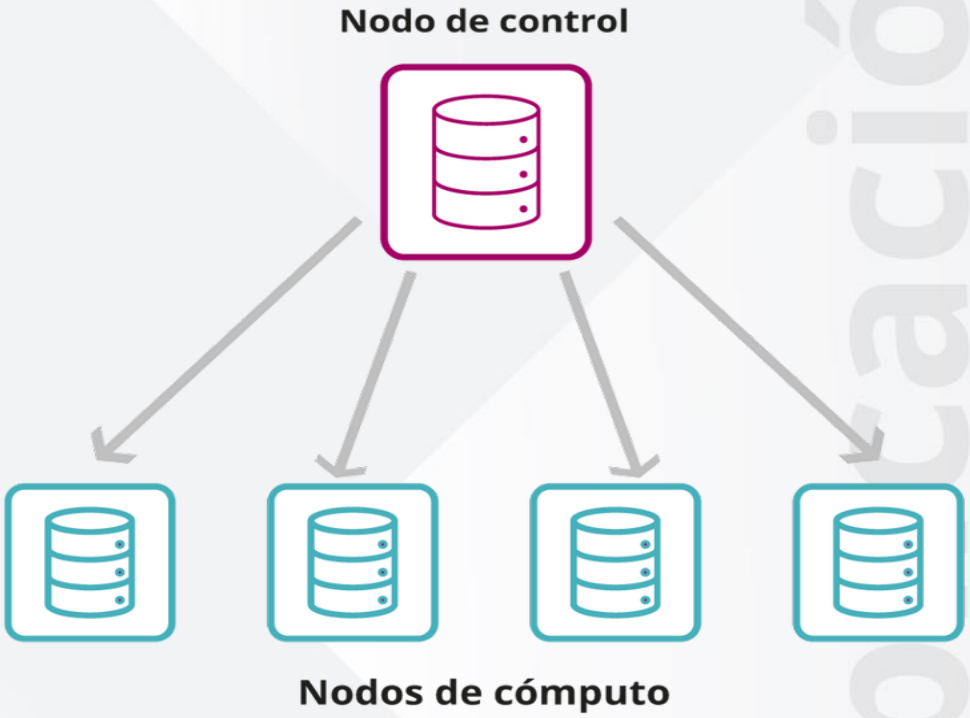
Distribuyen datos en un grupo de servidores, que permite que los sistemas compartan la carga de trabajo de procesamiento de consultas.

## Características

Procesamiento.
Interconexión de alta velocidad.
Bloqueo distribuido.

**Ejemplos**:
BigQuery
Snowflake
Synapse
Amazon Redshift

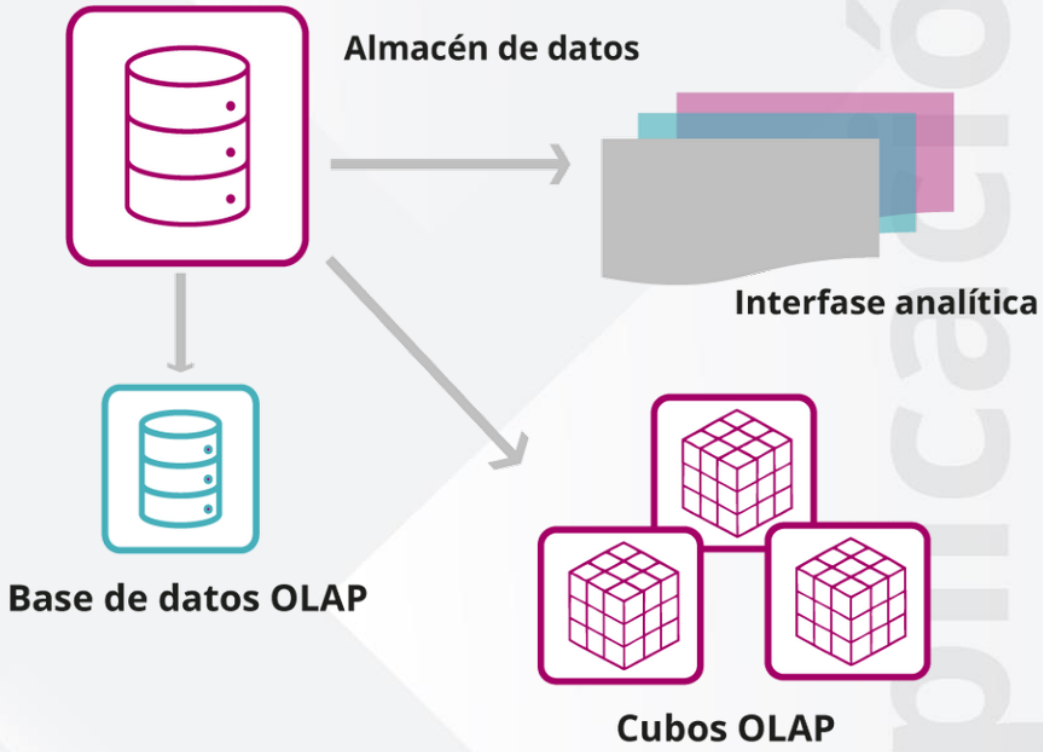
Figura 5. Ejemplo de base de datos de procesamiento paralelo masivo (MPP).




Fuente: Abdul, I. (2022). What are MPP systems? Benefits, types and examples. *Royal Cyber*. Recuperado de https://www.royalcyber.com/blog/data-services/what-is-massively-parallel-processing-mpp/

Universidad Tecmilenio

Figura 6. Bases de datos de procesamiento analítico en línea (OLAP).




Fuente: Altexsoft. (2021). *What is OLAP: A complete guide to online analytical processing*. Recuperado de https://www.altexsoft.com/blog/olap-online-analytical-processing/

Almacenan "cubos" multidimensionales de datos agregados para analizar información basada en múltiples atributos de datos.

**Ejemplos**:

- Apache Kylin
- Microsoft SQL Server Analysis Services (SSAS)
- IBM Cognos TM1

Universidad Tecmilenio

## Resumen

**Objetivo:** Resumir los tipos de bases de datos para analítica que existen en el mercado actual, sus características y empresas que las ofrecen.

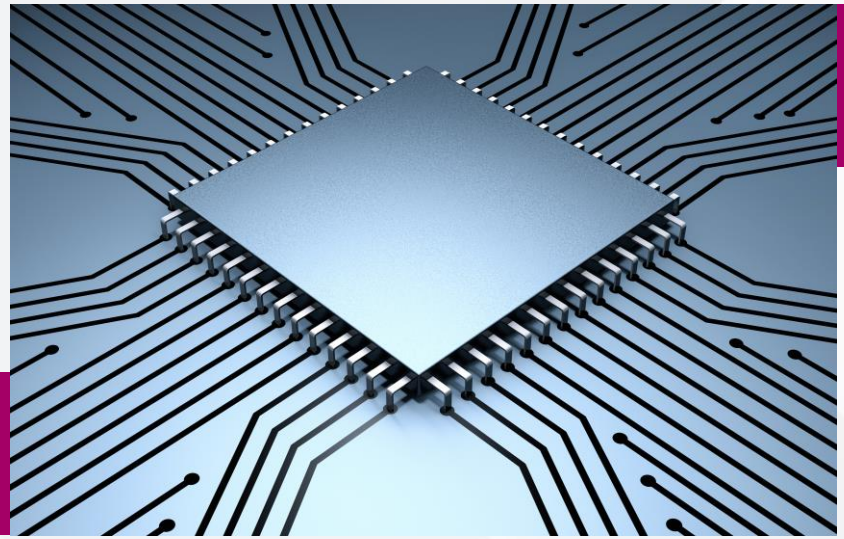
**Instrucciones:**

- Para realizar esta actividad es importante que consultes el contenido del tema.
- **Resume** los tipos de bases de datos para analítica que existen en el mercado actual, sus características y empresas que las ofrece
- Haz énfasis en los aspectos que mejor respondan a las necesidades de tu empresa.
- Recuerda presentar las fuentes de información aplicando los criterios APA e incluir colores y tipografías y material gráfico que apoyen visualmente tu trabajo.
- Máximo 1 cuartilla.

Universidad Tecmilenio

Se tiene una variedad de opciones en cuanto bases de datos de analítica se refiere, siendo esta elección uno de los puntos clave del departamento de **ciencias de datos,** en conjunto con la **arquitectura** y el **gobierno** de estos.

Las dimensiones de la empresa, de los datos y de lo que se quiere lograr con ellos son los **indicadores** que deben guiar esta decisión teniendo en cuenta toda la tecnología de la que ya se puede echar mano, así como de las experiencias previas.



# Referencias


Abdul, I. (2022). What are MPP systems? Benefits, types and examples. *Royal Cyber*. Recuperado de https://www.royalcyber.com/blog/data-services/what-is-massively-parallel-processing-mpp/

Altexsoft. (2021). *What is OLAP: A complete guide to online analytical processing*. Recuperado de https://www.altexsoft.com/blog/olap-online-analytical-processing/

DataChannel. (2021). *What is Data Warehousing? How it Works, Types, and General Stages*. Recuperado de https://datachannel.co/blogs/introduction-to-data-warehousing/

GridGain. (s.f.). *In-Memory Cache*. Recuperado de https://www.gridgain.com/resources/glossary/in-memory-computing-platform/in-memory-cache

Holistics. (2021). The analytics setup guidebook. Recuperado de https://www.holistics.io/books/setup-analytics/a-modern-analytics-stack/

Scylla. (2022). *Columnar Database*. Recuperado de https://www.scylladb.com/glossary/columnar-database/

Universidad Tecmilenio

Fundamentos de big data

# Tema 8.
# Analítica de datos.

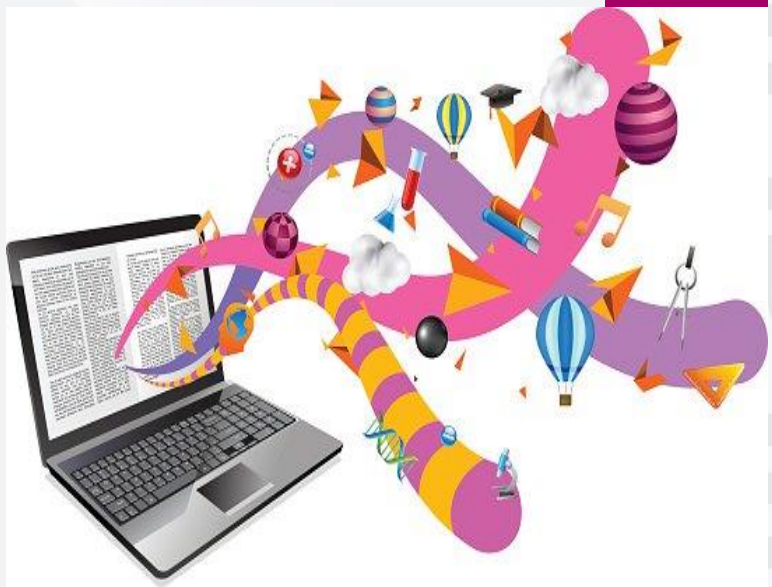
# Introducción

En esta época en donde todo esa cambiando constante y a veces rápidamente, donde como humanidad nos adecuamos a nuevas circunstancias, se están haciendo también nuevas iniciativas que nombramos disruptivas, porque son diferentes a lo establecido, en una época de más apertura, conocimiento y aceptación.

La **analítica de datos** no es solo parte de ello sino que contribuye a esta nueva sociedad y también en el cómo nos adaptarnos a ella.

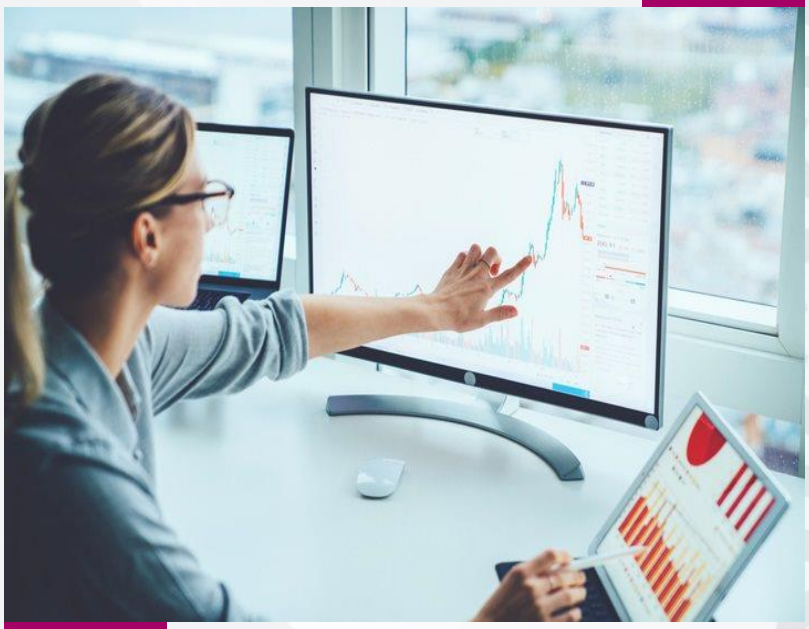
El trabajo del **analista de datos** es extraer los datos crudos, **organizarlos, analizarlos y transformarlos** a información coherente e inteligente.

**La Inteligencia comercial** es el principal beneficio de la analítica de datos,  ya que consiste en encontrar patrones en el conjunto de datos que puedan proporcionar información útil y relevante tanto para la organización como para un área específica de la misma.



Universidad Tecmilenio

Ayuda a **dar sentido a la información del pasado y al mismo tiempo predecir tendencias y comportamientos futuros** e incluso en la automatización de ciertas tareas dejando así de basar las decisiones y estrategias en conjeturas, y tomándolas, basándose en información y datos.



Explicación

Universidad Tecmilenio

## Diferencia entre analítica de datos y ciencia de datos

Un **analista** buscará responder preguntas específicas o abordar desafíos particulares que ya se han identificado y la empresa conoce.

Un **científico** de datos considera que preguntas debería o podría hacer la empresa, diseñan modelos predictivos y ejecutan análisis personalizados.

El **analista de datos** aborda y resuelve preguntas discretas sobre datos, generalmente solicitados, revelando información sobre la que pueden actuar las partes interesadas y los **científicos** crean sistemas para automatizar y optimizar el funcionamiento del negocio.

Figura 1. Diferenciación de herramientas y habilidades entre analista y científico de datos.


Fuente: Stevens, E. (2022-a). *What is data analytics? A complete guide for beginners*. Recuperado de https://www.sigmacomputing.com/blog/top-20-big-data-statistics/

Figura 2. Principales tipos de análisis de datos.


Fuente: Stevens, E. (2022-a). *What is data analytics? A complete guide for beginners*. Recuperado de https://www.sigmacomputing.com/blog/top-20-big-data-statistics/

**Análisis descriptivo:** Es un análisis simple a nivel superficial, que analiza lo que se sucedió en el pasado.

Las dos técnicas que se usan generalmente son: la agregación y la extracción de datos. Por lo que el analista recopila los datos primero y presenta un formato resumido (parte de la agregación) y luego examina los datos para encontrar los patrones.

**Análisis de diagnóstico:** El análisis de diagnóstico explora el "por qué". En este caso el analista, primero busca anomalías en los datos al final, intentará descubrir relaciones causales.

En esta etapa el analista puede apoyarse en la teoría de la probabilidad, el análisis de regresión, el filtrado y el análisis de datos de series temporales.

**Análisis. Predictivo:** Intenta predecir lo que sucederá en el futuro. Estima la probabilidad de un resultado futuro en función de los datos históricos y la teoría de la probabilidad

**Este análisis se pude usar para pronosticar todo tipo de resultados.**

**Análisis Prescriptivo:** Muestra cómo se pueden aprovechar los resultados que se han pronosticado. Es uno de los análisis más complejos que se pueden realizar y puede implicar trabajar con **algoritmos, aprendizaje automático y procedimientos de modelado computacional.**

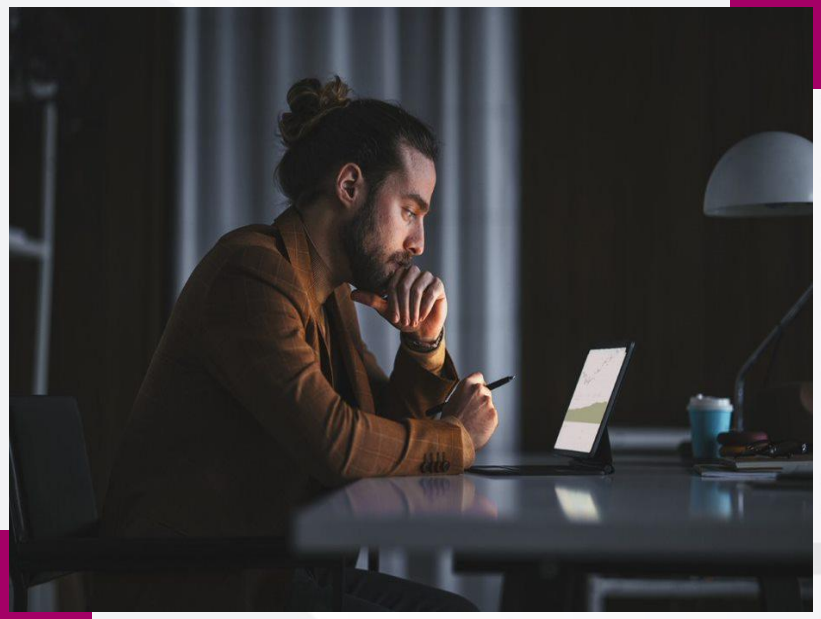
Universidad Tecmilenio

Una vez definido el tipo de análisis de datos hay que seguir un proceso para analizarlos.

Universidad Tecmilenio

**Los datos cuantitativos:** se refieren a cualquier cosa que sea medible.

**Los datos cualitativos:** por otro lado, no se pueden medir, existen proyectos que requieren de estos datos por lo que es importante considerarlos a la hora de recopilarlos y analizarlos.

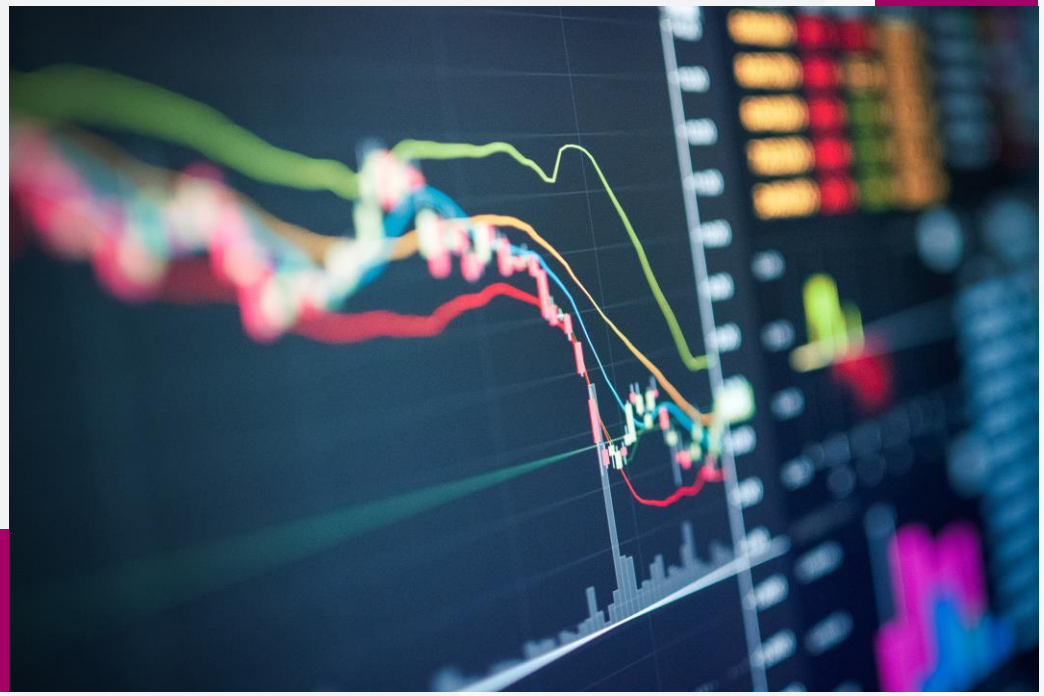


Explicación

Universidad Tecmilenio

Principales técnicas de análisis de datos:

- Análisis de regresión.
- Análisis factorial.
- Análisis de cohort o grupal.
- Análisis de conglomerados.
- Análisis de series de tiempo



Explicación

Universidad Tecmilenio

Existen otros tipos de análisis que también son comúnmente usados:

**Análisis de los sentimientos.** Es una técnica cualitativa, su objetivo es interpretar y clasificar las emociones transmitidas dentro de los datos en texto.

Esto le permite a la empresa determinar cómo se sienten sus clientes acerca de varios aspectos de su **marca, producto o servicio.**

**Simulación de Monte Carlo.** Es una técnica computarizada utilizada para generar modelos de posibles resultados y sus distribuciones de probabilidad. Esencialmente, considera un rango de resultados posibles y luego calcula la probabilidad de que se realice cada resultado en particular.

Explicación

Universidad Tecmilenio

**Secuencia de pasos:**

**Objetivo:** Elaborar una secuencia de pasos para el análisis de datos con fines predictivos que apoyen la elección de cambios en una marca cualquiera.

**Instrucciones:**

1. Para realizar esta actividad es importante que consultes el contenido del tema.
2. Elabora una **secuencia de pasos** para el análisis de datos con fines predictivos que apoyen la elección de cambios en una marca de tu preferencia, aplicando los conocimientos del tema.
3. En cada paso describe en qué consiste, las herramientas y/o fuentes a utilizar para su realización, incluye la presentación de la marca y el propósito del análisis.
4. Recuerda presentar las fuentes de información aplicando los criterios APA e incluir colores y tipografías y material gráfico que apoyen visualmente tu trabajo.
5. Máximo 1 cuartilla.

Universidad Tecmilenio

Al reconocer concretamente los análisis, el proceso y las técnicas más usadas de esta analítica actualmente nos deja ver más de cerca cómo puedes usarla y aportar para vislumbrar el futuro de ésta y el papel que quieras realizar en ella.

Cuando conoces el impacto que puede tener la analítica de datos en las organizaciones y en el mundo, puedes plantearte en qué quisieras aportar con el rol actual o el que decidas tomar.

# Referencias

Stevens, E. (2022-b). *The 7 Most Useful Data Analysis Methods and Techniques*. Recuperado de https://www.sigmacomputing.com/blog/top-20-big-data-statistics/